LEVERAGING EMERGING GENOMIC TECHNOLOGIES FOR THE BENEFIT OF THE DOMESTIC HORSE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Heather Marie Holl

May 2014

LEVERAGING EMERGING GENOMIC TECHNOLOGIES FOR THE BENEFIT OF THE DOMESTIC HORSE

Heather Marie Holl, Ph. D.

Cornell University 2014

The domestic horse has played an important role in human history and serves as a valuable model organism for genetic studies. Generations of selective breeding have created several key population and genomic advantages for study. The completion of the horse reference genome sequence in 2009 further strengthened the utility of the horse as a genetic model. However, in order to continue deriving the maximum benefit from research, it is vital to incorporate novel genomic technologies. We applied three such methodologies in the following projects.

Chromosomal aberrations can lead to congenital abnormalities, embryonic loss, and infertility. Detection of small deviations - such as aneuploidy in the acrocentric chromosomes, mosaicisms, or partial chromosomal structural variation - can be difficult with typical cytogenetics techniques. The Illumina Equine SNP50 BeadChip was utilized for molecular karyotyping, assessing copy number variation at fifty thousand loci simultaneously throughout the genome. This method successfully diagnosed aberrations in three cases where the traditional FISH was not possible.

Laminitis is a devastating disease of the hoof that is the second leading cause of both lameness and euthanasia in the horse. The etiology is poorly understood, and published research has focused on a small number of known gene targets. Due to the poor quality of current equine gene annotation, RNA-seq of lamellar tissue was

undertaken to provide a more robust set of targets. *De novo* transcriptome assembly generated a valuable gene annotation resource for the future study of laminitis.

Leopard complex spotting is a unique collection of pigmentation patterns in which a single autosomal allele (*LP*) allows for the expression of other modifying loci. RNA-seq was used to supplement typical gene mapping approaches for characterizing *PATN1*, a major effect modifier of *LP*. A targeted set of variants was produced for fine-mapping a 16 Mbp region associated with the phenotype. These genotypes identified a SNP in the 3'UTR of *RFWD3* that may be utilized by horse breeders for the identification of non-*LP* horses that carry *PATN1*.

These three projects demonstrate the benefit of next-generation technologies. High-throughput methods facilitate rapid high-resolution evaluation of the entire genome, providing a comprehensive tool for future research.

BIOGRAPHICAL SKETCH

Heather Marie Holl was born on December 7th, 1988 in Silver Spring, Maryland. She graduated from Brunswick High School in 2004. She attended Otterbein College until 2009, where she graduated *cum laude* with honors with a BSc, dual majoring in Equine Science Pre-Veterinary/Pre-Graduate Studies and Life Science Molecular Biology. During her BSc, she worked under Dr. Simon Lawrance studying the genetic basis of the heritable brindle coat color in horses. In 2009, she entered the PhD program in Animal Science at Cornell University under the supervision of Dr. Samantha Brooks, working on incorporating emerging genomic technologies in the domestic horse.

To my family,

ACKNOWLEDGMENTS

I would like to thank all of the educators who have guided me along the way, providing me with a life-long love of learning. My special committee members have been an important source of assistance during my degree, and I have thoroughly enjoyed the classes, collaborations, and meetings I have had with them. I am especially grateful to Dr. Samantha Brooks for her guidance and support throughout my degree, allowing me to combine my interests of computers, genetics, and horses in ways that I never imagined were possible. Special thanks as well go to Drs. Maria Calderone and Simon Lawrance, whose mentoring during my BSc helped me transition from planning for veterinary school to instead pursuing a PhD, and who introduced me to Dr. Ernie Bailey. Dr. Bailey also deserves special mention, for his encouragement throughout my travels, as well as for our adventures during the various conferences I have attended. I would like to thank Dr. Rebecca Bellone for her support and for welcoming me onto the Appaloosa Project, providing me with the opportunity to study one of my favorite horse coat colors.

Being a member of the horse genetics research community has been a wonderful experience. I have thoroughly enjoyed all of the conferences I have attended, allowing me to meet many brilliant researchers from around the world. To this end, I am grateful for the travel funding I have received from the Cornell University Graduate School and the USDA NAGRP NRSP-8 program,which have allowed me to venture beyond Cornell's campus.

I would also like to thank the crew up on the fourth floor in Morrison for their friendship and tolerance of my many quirks. Special mentions go to Ann Staiger, who I can say I've traveled the world with, and to Fernando Migone, who has been one of my most supportive friends. Additionally, I thank the regular group of graduate

students that hang around after our Thursday seminars, who have provided valuable stress-relieving conversations, especially during this final semester. To our final two undergraduate researchers as well, Lauren Jones and Chris Posbergh, thank you for your willingness to help whenever possible.

The Department of Animal Science has been a truly wonderful place to do my PhD. I am forever grateful to the funding that has allowed me to continue with my studies, and to all of the faculty and staff that I have had the pleasure to meet and work with. I will never forget my time here in Morrison Hall.

Outside of the world of academia, I am of course grateful to my family and all of my friends. Despite not always quite understanding what I am excited about or why, they have been wonderfully supportive of my studies. I may not have been able to pull myself away to visit very often, but I have always kept you all in my thoughts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The horse has played a vital role in human history, assisting in the spread of culture and exploration of new lands (Anthony 2010). Although generally no longer required as a work animal, the domestic horse can serve as an important model system for genetics research. There are more than 200 hereditary equine conditions reported in the scientific literature, many of which display high similarity to human diseases (OMIA, Chowdhary *et al*. 2008). The mating systems used in modern breeds have also led to several beneficial genome features. Many haplotypes are shared across breeds and display fairly moderate linkage disequilibrium, providing a genetic landscape more similar to humans than that of dogs or mice (Wade *et al*. 2009). Similarly, the horse genome (comprised of 31 pairs of autosomes and the sex chromosomes) is highly syntenic with the human genome, with the majority of equine chromosomes corresponding to sections from just one or two human chromosomes (Yang *et al*. 2004). Finally, there is an active community of researchers dedicated to furthering equine genetics and a thriving industry ready to capitalize on associated discoveries.

Initially, equine genetic research was conducted primarily in individual laboratories, each focused on mapping single chromosomes. In 1995, a collaborative genome project was officially formed in order to develop a complete genetic map (Guérin *et al*. 1999). A variety of maps have been constructed and updated since then, including BAC contig, cytogenetic, linkage, radiation hybrid (RH), and Y chromosome maps, as well as comparative maps to other species (reviewed in Chowdhary and Raudsepp 2008). One of the latest maps to be produced integrated markers from all of the previous types, resulting in a set of 4,103 genome-wide markers (Raudsepp *et al*. 2008). The integrated map included 3,816 RH and 1,144

FISH markers, as well as 920 linkage and 1,904 comparative markers, yielding an average spacing of 775 kbp. Even as newer tools have been developed (as described below), this resource is still in common use today.

Genetic mapping involves the statistical testing of the correlation between a phenotype and polymorphic markers, either through linkage (within families) or association (among unrelated individuals). Historically, microsatellites have been the markers of choice largely due to their multiallelic nature (Gulcher 2012). Unlike single nucleotide polymorphisms (SNPs, which usually only have two alleles at a given loci), microsatellites have a higher information content from genotyping. As the mutation rate is higher for microsatellites, including these markers creates more haplotype diversity and shorter blocks of linkage disequilibrium, improving map resolution. However, while SNP genotyping can easily be multiplexed on large arrays to examine hundreds of thousands of loci at once, microsatellite genotyping is more difficult to perform, and even so-called high-density scans are still on the order of hundreds to thousands of markers.

Despite this, several fairly recent horse studies have incorporated microsatellite data, even using these markers alongside the newer equine SNP genotyping arrays (Andersson *et al*. 2011, Brault *et al*. 2011, Fox-Clipsham *et al*. 2011, Shakhsi-Niaei *et al*. 2011). In Chapter Four, targeted microsatellite linkage mapping was used to localize our phenotype of interest on ECA3, which bypassed the need for the more expensive genome-wide SNP chip. In addition to genetic mapping, the high allelic variability makes these markers ideal for population genetics studies and parentage verification, ensuring the use of microsatellites for years to come (Gulcher 2012, Bowling *et al*. 1997).

The integrated marker map is also applied in detecting chromosome structure through karyotyping and fluorescent *in situ* hybridization (FISH). In terms of

hereditary phenotypes, the horse is fairly unique in that a large paracentric chromosomal inversion is responsible for a white spotting pattern, without any apparent reduction in fitness or fertility (Brooks *et al*. 2007). Though a PCR-based test has since been developed for the purpose of genotyping, the inversion was initially characterized using FISH. Yet similar as to in humans, most known chromosomal abnormalities in the horse present as sub- or infertility, or as congenital morphological defects in foals (Lear and Bailey 2008). The majority of cytogenetics analysis in domestic animals is currently utilized for diagnostics in clinical cases (Ducos *et al*. 2008). Chromosome banding patterns are often insufficient to detect chromosomal translocations or to differentiate the smaller acrocentric chromosomes, though FISH can resolve these structures using markers from the physical map and reference genome. However, in Chapter Two an alternate cytogenetics technique is presented, in which the genotyping array was used to detect chromosomal aberrations where traditional cytogenetic techniques were not practical.

The Broad Institute completed assembly of the equine genome sequence (EquCab2) in 2009. EquCab2 is a high-quality assembly comprised of whole genome shotgun Sanger reads to approximately a 6.8x depth of coverage. Assembly yielded a N50 size (value at which 50% of assembled sequences are greater than or equal to) of 112 kbp for contigs and 46 Mbp for scaffolds (Wade *et al*. 2009). Over 95% of assembled sequences matched to the existing physical map. Along with the reference animal, a Thoroughbred named Twilight, an additional seven individual horses from diverse breeds were sequenced at low coverage in order to produce a database of genetic polymorphisms. The identified SNPs were used in the design of a 50K marker Illumina Equine SNP50 BeadChip (mentioned above as the genotyping array), which was developed and released in 2008 (McCue *et al*. 2012). After the SNP50 was discontinued, a 70K Illumina array (Equine SNP70) was produced, and design of an

Affymetrix 685K array is currently underway. SNP array technology has proven useful in the study of genetics of disease and performance in the horse and will likely continue to have a place in future research (Brooks and Bailey 2013).

While the genome assembly is high-quality, it is not perfect. At the time of publication, there were 9,604 contigs that could not be placed (N50 = 52,972), accounting for 93 Mbp of sequence or 5% of the genome ("chromosome unknown," chrUn). However, beyond the known problematic chrUn, a study examining linkage between markers on the initial genotyping array identified potential regions of mis-assembly (Corbin *et al*. 2012). In two cases, clusters of SNPs were identified that had no linkage to their reported neighbors, but instead were closely associated to markers on a different chromosome. Furthermore, 39 SNPs were closely linked with markers over 10 Mbp away. Some of these discordant placements may represent genuine structural variations between mapped breeds and the reference animal. However, most are likely erroneous placements during computational assembly. Additional sequencing and mapping by FISH is needed to verify the assembly in certain regions of the reference genome.

In addition to sequence errors, gene annotation is also problematic. Currently, most gene data is derived from computational predictions and cross-species annotation. One study was able to identify 227 expressed protein-coding genes that had annotation in other species, but were completely absent from the equine gene set (MacLeod *et al*. 2013). The equine genomics community is currently involved in the creation of a third version of the assembly, using advanced sequencing technologies to improve the current sequence and provide sufficient data for a more robust gene annotation set (Kalbfleisch *et al*. 2014).

A reference genome assembly also enables study of another important source of variation: copy number variants (CNVs). CNVs are segments of the genome that

have been lost or duplicated in certain individuals. Though many phenotypes have been linked to CNVs in domestic animals, so far only the grey coat color has been described in the horse (Clop *et al*. 2012, Rosengren-Pielberg *et al*. 2008). However, once the CNV content of the genome is better characterized, it is likely that additional phenotypes will be attributed to this type of polymorphism.

Equine Illumina genotyping arrays identified CNVs in two separate studies (Dupuis *et al*. 2013, Metzger *et al*. 2013). A similar method was also used in Chapter Two of this dissertation to detect a large scale CNV. However, while the planned 600K Affymetrix array is a vast improvement in the number of genotyped markers, it does not have the ability to assess DNA quantity. Two custom comparative genomic hybridization arrays exist for this purpose, though so far no association studies have utilized them (Doan *et al*. 2012a, Wang *et al*. 2014). A third method for detecting CNVs is through whole genome re-sequencing and quantification of the resulting read depth. This method was recently applied in a short-read resequenced genome from a Quarter Horse (Doan *et al*. 2012b). Next-generation sequencing technologies such as this may be the method of choice for future studies.

The resequencing of the Quarter Horse genome represented an important milestone in equine genomics. The initial horse genome project was the result of years of planning, months of sequencing, a large effort of many research groups, and at a cost of approximately 25 million dollars. However, with the advent of so-called next generation sequencing, researchers gained the ability to generate sufficient data to sequence a genome less than a week with costs more in the range of thousands of dollars. A review of the advances of technologies and details of each is presented by Liu *et al*. 2012. The methods most relevant to this dissertation are Sanger and Illumina sequencing. Developed in 1877, Sanger sequencing, also known as first generation sequencing, involves chain termination chemistry on a sample of many random length

DNA fragments followed by capillary sequencing and fluorescence detection. This allows for single base resolution at high confidence for long sequences (up to 1,000 bases), but at high expense and long run times. This technology was used for the initial horse reference genome assembly, and is the method utilized in Chapters Three and Four for validation of reported sequences.

The next generation of sequencers were developed in the mid-2000s, with the Illumina system released in 2006. Now one of the most popular methods, Illumina sequencing involves a random sheering of DNA fragments followed by clonal amplification to form a library. Fluorescence detection during step-by-step DNA synthesis identifies which base is incorporated at each cycle for millions of molecules simultaneously. This technology produces much shorter sequences than Sanger (25 to 250 bases) with a higher error rate (around 2%), but has the ability to produce hundreds of billions of bases in a week long run. Due to the high depth of coverage produced, the impact of the error rate is easily reduced. Illumina sequencing is currently one of the cheapest of sequencing technologies, recently boasting the ability to generate an entire human genome for less than $1,000 USD with the HiSeq X Ten.

One application of next-generation sequencing is whole transcriptome sequencing (RNA-seq). As the name indicates, total RNA is subjected to shotgun sequencing methods, which produces a measure of the transcriptome without the need for prior knowledge of sequences. This is especially ideal for situations where annotation is not available, whether it be for assessing differential expression in a novel tissue type, or for expression or variant discovery studies in non-model organisms. For analysis, there are generally two methods to handle RNA-seq data (Martin and Wang 2011). The first option is to map sequences to a reference genome, which is computationally efficient and highly sensitive. However, as this method is limited by the quality and availability of the reference genome, it may not be ideal for

all species, and may not accurately detect exon splicing. Alternatively, the sequencing reads can be *de novo* assembled into transcripts representing the original mRNAs, which avoids the problems of the reference-based approach. Nevertheless, *de novo* assembly is not without its own issues. The repetitive nature of mammalian genomes often leads to shorter sequences being assembled incorrectly, and the algorithms to perform these assemblies require significant computing power. These two methods can also be combined, leveraging the advantages of both methods to provide a more complete view of the transcriptome. In Chapter Three, *de novo* assembly is performed in a specialized equine tissue, with reference-based approaches used to generate variant calls and gene annotation for the current EquCab2 reference genome. RNA-seq also appears in Chapter Four, first to generate variants to create fine-mapping panels, but also utilizing the *de novo* assembly from Chapter Three to further analyze candidate genes.

RNA-seq was used previously to successfully map leopard complex spotting (*LP*, Bellone *et al*. 2013). This coat color, reviewed in Chapter Four, had been mapped to ECA1 using microsatellites and associated with reduced expression of *TRPM1*, but sequencing failed to identify a causative mutation. Three strongly associated SNPs were identified within introns, so transcriptome sequencing was attempted in order to determine if these SNPs fell within novel expressed or coding regions. While the regions containing these SNPs were not expressed, a region of abnormal intronic expression was observed, and further molecular work identified a retroviral insertion with perfect association to *LP*. This insertion results in premature polyadenylation of *TRPM1*, suggesting it was the causative variant. Chapter Four presents the continuation of this work, applying a similar methodology to mapping a major modifier of the *LP* phenotype.

As overviewed above, incorporation of emerging analyses and technologies is vital for the future of equine genomics. The studies in this dissertation thus help move the field forward, providing important tools for the future.

REFERENCES

Andersson LS, Lyberg K, Cothran G, Ramsey DT, Juras R, Mikko S, Ekesten B, Ewart S, Lindgren G (2011) Targeted analysis of four breeds narrows equine Multiple Congenital Ocular Anomalies locus to 208 kilobases. Mamm Genome 22: 353-360.

Anthony DW (2007) The Horse, the Wheel, and Language. Princeton Univ. Press, Princeton, NJ, USA.

Bellone RR, Holl H, Setaluri V, Devi S, Maddodi N, Archer S, Sandmeyer L, Ludwig A, Foerster D, Pruvost M, Reissmann M, Bortfeldt R, Adelson DL, Lim SL, Nelson J, Haase B, Engensteiner M, Leeb T, Forsyth G, Mienaltowski MJ, Mahadevan P, Hofreiter M, Paijmans JL, Gonzalez-Fortes G, Grahn B, Brooks SA (2013) Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. PLoS One 8: e78280.

Bowling AT, Eggleston-Stott ML, Byrns G, Clark RS, Dileanis S, Wictum E (1997) Validation of microsatellite markers for routine horse parentage testing. Anim Genet 28: 247-252.

Brault LS, Cooper CA, Famula TR, Murray JD, Penedo MC (2011) Mapping of equine cerebellar abiotrophy to ECA2 and identification of a potential causative mutation affecting expression of MUTYH. Genomics 97: 121-129.

Brooks SA, Bailey E (2013) Horse Genetics. Second Edition. CABI, Boston, MA, USA.

Brooks SA, Lear TL, Adelson DL, Bailey E (2007) A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. Cytogenet Genome Res 119: 225-230.

Chowdhary BP, Paria N, Raudsepp T (2008) Potential applications of equine genomics in dissecting diseases and fertility. Anim Reprod Sci 107: 208-218.

Chowdhary BP, Raudsepp T (2008) The horse genome derby: racing from map to whole genome sequence. Chromosome Res 16: 109-127.

Clop A, Vidal O, Amills M (2012) Copy number variation in the genomes of domestic animals. Anim Genet 43: 503-517.

Corbin LJ, Blott SC, Swinburne JE, Vaudin M, Bishop SC, Woolliams JA (2012) The identification of SNPs with indeterminate positions using the Equine SNP50 BeadChip. Anim Genet 43: 337-339.

Doan R, Cohen N, Harrington J, Veazey K, Juras R, Cothran G, McCue ME, Skow L, Dindot SV (2012) Identification of copy number variants in horses. Genome Res 22: 899-907.

Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV (2012) Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. BMC Genomics 13: 78.

Ducos A, Revay T, Kovacs A, Hidas A, Pinton A, Bonnet-Garnier A, Molteni L, Slota E, Switonski M, Arruga MV, van Haeringen WA, Nicolae I, Chaves R, Guedes-Pinto H, Andersson M, Iannuzzi L (2008) Cytogenetic screening of livestock populations in Europe: an overview. Cytogenet Genome Res 120: 26-41.

Dupuis MC, Zhang Z, Durkin K, Charlier C, Lekeux P, Georges M (2013) Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy. Anim Genet 44: 206-208.

Fox-Clipsham LY, Carter SD, Goodhead I, Hall N, Knottenbelt DC, May PD, Ollier WE, Swinburne JE (2011) Identification of a mutation associated with fatal Foal Immunodeficiency Syndrome in the Fell and Dales pony. PLoS Genet 7: e1002133.

Guérin G, Bailey E, Bernoco D, Anderson I, Antczak DF, Bell K, Binns MM, Bowling AT, Brandon R, Cholewinski G, Cothran EG, Ellegren H, Förster M, Godard S, Horin P, Ketchum M, Lindgren G, McPartlan H, Mériaux JC, Mickelson JR, Millon LV, Murray J, Neau A, Røed K, Ziegle J, et al (1999) Report of the International Equine Gene Mapping Workshop: male linkage map. Anim Genet 30: 341-354.

Guérin G, Bailey E, Bernoco D, Anderson I, Antczak DF, Bell K, Biros I, Bjørnstad G, Bowling AT, Brandon R, Caetano AR, Cholewinski G, Colling D, Eggleston M, Ellis N, Flynn J, Gralak B, Hasegawa T, Ketchum M, Lindgren G, Lyons LA, Millon LV, Mariat D, Murray J, Neau A, Røed K, Sandberg K, Skow LC, Tammen I, Tozaki T, Van Dyk E, Weiss B, Young A, Ziegle J (2003) The second generation of the International Equine Gene Mapping Workshop half-sibling linkage map. Anim Genet 34: 161-168.

Gulcher J (2012) Microsatellite markers for linkage and association studies. Cold Spring Harb Protoc 2012: 425-432.

Kalbfleisch T, Rebolledo-Mendez J, Orlando L, MacLeod JN (2014) Resources, and Progress Toward a Fully Annotated EquCab3. Plant and Animal Genome XXII. January 12th, San Diego, California, USA

Lear TL, Bailey E (2008) Equine clinical cytogenetics: the past and future. Cytogenet Genome Res 120: 42-49.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechno 2012: 251364.

MacLeod JN, Zeng Z, Hestand M, Coleman S, Orlando L, Kalbfleisch T (2013) Annotated protein-coding genes that are missing from EquCab2. 10th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop. July 11th, Furnas, Azores, Portugal.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12: 671-682.

Metzger J, Philipp U, Lopes MS, da Camara Machado A, Felicetti M, Silvestrelli M, Distl O (2013) Analysis of copy number variants by three detection algorithms and their association with body size in horses. BMC Genomics 14: 487.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Distl O, Guérin G, Hasegawa T, Hill EW, Leeb T, Lindgren G, Penedo MC, Røed KH, Ryder OA, Swinburne JE, Tozaki T, Valberg SJ, Vaudin M, Lindblad-Toh K, Wade CM, Mickelson JR (2012) A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. PLoS Genet 8:ve1002451.

OMIA (2009) Online Mendelian Inheritance in Animals. University of Sydney, http://omia.angis.org.au/.

Penedo MC, Millon LV, Bernoco D, Bailey E, Binns M, Cholewinski G, Ellis N, Flynn J, Gralak B, Guthrie A, Hasegawa T, Lindgren G, Lyons LA, Røed KH, Swinburne JE, Tozaki T (2005) International Equine Gene Mapping Workshop Report: a comprehensive linkage map constructed with data from new markers and by merging four mapping resources. Cytogenet Genome Res 111: 5-15.

Raudsepp T, Gustafson-Seabury A, Durkin K, Wagner ML, Goh G, Seabury CM, Brinkmeyer-Langford C, Lee EJ, Agarwala R, Stallknecht-Rice E, Schäffer AA, Skow LC, Tozaki T, Yasue H, Penedo MC, Lyons LA, Khazanehdari KA, Binns MM, MacLeod JN, Distl O, Guérin G, Leeb T, Mickelson JR, Chowdhary BP (2008) A 4,103 marker integrated physical and comparative map of the horse genome. Cytogenet Genome Res 122: 28-36.

Rosengren Pielberg G, Golovko A, Sundström E, Curik I, Lennartsson J, Seltenhammer MH, Druml T, Binns M, Fitzsimmons C, Lindgren G, Sandberg

K, Baumung R, Vetterlein M, Strömberg S, Grabherr M, Wade C, Lindblad-Toh K, Pontén F, Heldin CH, Sölkner J, Andersson L (2008) A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. Nat Genet 40: 1004-1009.

Shakhsi-Niaei M, Klukowska-Rötzler J, Drögemüller C, Swinburne J, Ehrmann C, Saftic D, Ramseyer A, Gerber V, Dolf G, Leeb T (2011) Replication and fine-mapping of a QTL for recurrent airway obstruction in European Warmblood horses. Anim Genet 43: 627-631.

Swinburne J, Gerstenberg C, Breen M, Aldridge V, Lockhart L, Marti E, Antczak D, Eggleston-Stott M, Bailey E, Mickelson J, Røed K, Lindgren G, von Haeringen W, Guérin G, Bjarnason J, Allen T, Binns M (2000) First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families. Genomics 66: 123-134.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867.

Wang W, Wang S, Hou C, Xing Y, Cao J, Wu K, Liu C, Zhang D, Zhang L, Zhang Y, Zhou H (2014) Genome-wide detection of copy number variations among diverse horse breeds by array CGH. PLoS One 9: e86860.

Yang F, Fu B, O'Brien PC, Nie W, Ryder OA, Ferguson-Smith MA (2004) Refined genome-wide comparative map of the domestic horse, donkey and human based on cross-species chromosome painting: insight into the occasional fertility of mules. Chromosome Res 12: 65-76.

CHAPTER 2

DETECTION OF TWO EQUINE TRISOMIES USING SNP-CGH

H. M. Holl[1], T. L. Lear[2], R. D. Nolen-Walston[3], J. Slack[3], and S. A. Brooks[4*]

Published as an article in Mammalian Genome *(2013) 24(5-6): 252-256.*

[1]Department of Animal Science, Cornell University, Ithaca, NY 14853, USA

[2]Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA

[3]New Bolton Center, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

[*]Corresponding author

ABSTRACT

Chromosomal aberrations in many species, including the horse, are known to cause congenital abnormalities, embryonic loss, and infertility. While diagnosed mainly by karyotyping and FISH in the horse, the use of SNP array comparative genome hybridization (SNP-CGH) is becoming increasingly common in human diagnostics. Normalized probe intensities and allelic ratios detect changes in copy number genome-wide. Two horses with suspected chromosomal abnormalities and six horses with FISH-confirmed aberrant karyotypes were chosen for genotyping on the Equine SNP50 array (Illumina, Inc). Karyotyping of the first horse indicated mosaicism for an additional small, acrocentric chromosome, although the identity of the chromosome was unclear. The second case displayed a similar phenotype to human disease caused by a gene deletion, and so was chosen for SNP-CGH due to the ability to detect changes at higher resolutions than those achieved with conventional karyotyping. The results of SNP-CGH analysis for the six horses with known chromosomal aberrations agreed completely with previous karyotype and FISH analysis. The first undiagnosed case showed a pattern of altered allelic ratios without a noticeable shift in overall intensity for chromosome 27, consistent with a mosaic trisomy. The second case displayed a more drastic change in both values for chromosome 30, consistent with a complete trisomy. These results indicate that SNP-CGH is a viable method for detection of chromosomal aneuploidies in the horse.

INTRODUCTION

Chromosomal aberrations in the horse are known to cause congenital abnormalities, embryonic loss, and infertility (Lear and Bailey 2008). Cases are routinely diagnosed using chromosome banding techniques and fluorescent *in situ* hybridization (FISH). However, these techniques tend to be time consuming and require cultured cells. In human clinical diagnostics, the use of SNP array comparative genome hybridization (SNP-CGH) to obtain high resolution copy number estimates is gaining in popularity (Faas *et al.* 2012, Srebniak *et al.* 2012).

SNP-CGH is a robust analysis that uses fluorescence intensity values and allelic ratios from genotyping arrays to obtain high-resolution copy number estimates. Although resolution is dependent on the depth of markers available on the chip, even smaller scale arrays offer much higher resolution than other methods. It has the ability to detect amplifications, deletions, duplications, and copy-neutral loss of heterozygosity (Peiffer *et al.* 2006). Additionally, SNP-CGH is able to detect chimerism and mosaicism with around 5-20% abnormal cells in a sample (Conlin *et al.* 2010). Genotypes can be obtained from 50 ng of DNA without the need to culture cells (Srebiak *et al.* 2012).

The Equine SNP50 BeadChip uses Illumina Infinium II chemistry to genotype 54,602 loci throughout the genome with an average spacing of 43.1 kb. Cluster files indicating expected intensities were obtained from a training set of 354 horses representing 14 breeds (McCue *et al.* 2012). These cluster files are used to infer genotypes and as a comparison for SNP-CGH.

Two horses with suspected chromosomal aberrations were selected for SNP-CGH analysis. Both horses were admitted to a referral hospital due to congenital physiological abnormalities. As a proof of concept, six additional horses that had previously been karyotyped with FISH were included as well.

MATERIALS AND METHODS

Sample collection and case histories

Case one was a six month old Friesian filly with congenital skeletal abnormalities, failure to thrive, a brindle haircoat pattern, and abnormal external genitalia (Figure 2.1). A sample of blood along with a skin biopsy from the brindled region were submitted for genetic screening. At two years of age, the filly was euthanized and additional samples of skin and ovarian tissue were collected.

Case two was a four year old Welsh Pony colt with a ventricular septal defect, pulmonary artery hypoplasia, scoliosis, and a facial deformity. Due to phenotypic similarities to human DiGeorge syndrome, characterized by a large chromosomal deletion, blood was submitted for genetic testing.

An additional six horses with FISH-confirmed aberrant karyotypes were selected for SNP-CGH analysis to be used as controls. There were three 63,XO samples, one 65,XX+31 sample, one 65,XY+31 sample, and one 64,XX,t(1;21) sample (Table 2.1).

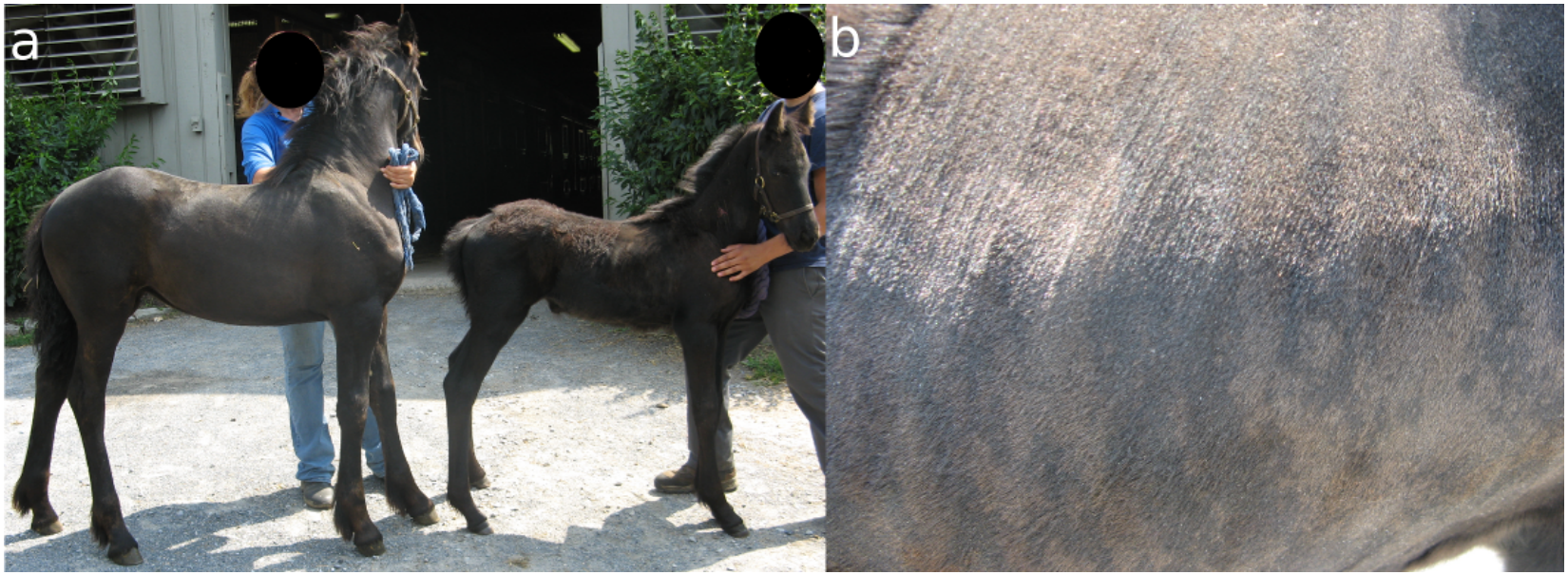

Figure 2.1. Outward appearance of case one at six months of age. **a** The filly (right) is shown next to a normal six month old Friesian weanling (left). **b** When fur from the side of the filly was clipped off, a brindled pattern of two colors was apparent.

Cell culture and FISH

Whole blood from case one was collected in sodium heparin and cultured as described previously (Lear *et al.*, 1999). Briefly, lymphocytes were cultured in RPMI-1640, supplemented with 10% fetal bovine serum, L-glutamine, antibiotic-antimycotic, phorbol/lectin or lectin (all Invitrogen). After 70 hours of culture, the cells were treated for 25 minutes in colcemid followed by 25 minutes in 0.067 M KCL. After a prefix with 3:1 methanol:acetic acid, the fixative was changed three or four times.

Tissues derived from a sample of skin and ovary from case one were digested in collagenase for 4 hours. The cells were cultured in FGM2: MEMα (50:50) (Clonetics) for several weeks to establish a fibroblast cell line on each tissue. Excess tissues and cells were archived in the Frozen Zoo® (Lab#18571).

Slides for FISH and karyotyping were prepared from the cell cultures. Colcemid was not added at harvest. Cells were treated for 30 minutes in 0.067 M KCl and fixed as usual. Slides were prepared using standard cytogenetic techniques. Chromosomes were GTG-banded by a modified method of Seabright (1971) and karyotyped according to the current karyotype standard for the horse (ISCNH, 1997).

Horse chromosome-specific BAC clones representing ECA27 (CH241-151K03) and ECA30 (CH241-139M13) were labeled with either Spectrum Orange or Green (Abbott Molecular) and hybridized to metaphase and interphase cells as described previously (Lear *et al.*, 2001). The cells were then scored according to the number of signals observed.

SNP genotyping

DNA was isolated from each of the tissues samples using the Qiagen Gentra Puregene Blood Kit (Qiagen Sciences, Germantown, MD, USA). The blood samples

were extracted using the “DNA Purification from Whole Blood” protocol whereas all other samples were extracted with the “DNA Purification from Mouse Tail Tissue” protocol. 1000 ng of DNA from the center of the first skin biopsy from case one and 1000 ng of DNA from blood from case two and the six controls were sent to Geneseek for genotyping on the Equine SNP50 BeadChip (Geneseek, Lincoln, NE, USA).

SNP-CGH analysis

The raw intensity files for all samples were obtained and entered into GenomeStudio for preliminary analysis (Illumina Inc., San Diego, CA, USA). Standard equine cluster files were used to generate the log2R ratios (log-transformed ratio of observed probe intensities compared to expected intensities from a training set, LRR) and B allele frequencies (estimate of allele frequency based on ratio of observed intensities for each allele probe, BAF) for each SNP. For a normal diploid cell, each allele of a SNP should be present in equal amounts, resulting in an average LRR of 0 and three clusters of BAF (AA homozygotes around 0, AB heterozygotes around 0.5, and BB homozygotes around 1). Loss of one chromosome will cause a decrease in LRR and elimination of the heterozygous BAF cluster, whereas a copy gain will increase LRR and create a split in BAF (AAB heterozygotes around 0.3 and ABB heterozygotes around 0.7). Calculation of LRR and BAF, as well as determination of a copy loss or gain, is described in detail in Peiffer *et al.* 2006. These values were then imported into R and first visualized on the genome scale for each sample individually. Abnormal regions were then identified by visual inspection on the chromosomal scale (R Development Core Team 2009).

Four additional values were calculated in R for visually abnormal chromosomes. For both monosomies and trisomies, the LRR of all SNPs on the affected chromosome were averaged (“Average LRR”). In the case of monosomies,

the total number of SNPs in between a BAF of 0.2 and 0.8 (representing possible heterozygous SNPs, “Numeric BAF”) were counted. Finally, for chromosomes with two central clusters of BAF (representing possible trisomic AAB and ABB heterozygous SNPs), two averages were calculated: average BAF of values from 0.2 to 0.5 and average BAF of values from 0.5 to 0.8 (“Numeric BAF”). Each average was compared to the corresponding averages from the control.

RESULTS

General SNP-CGH findings

On initial examination, large runs of homozygosity were noticed for almost every sample across the genome. Female individuals were noted to have an elevated LRR for the X chromosome. In all previously diagnosed horses, visual examination and calculated values agreed with the FISH karyotype. Full results for each sample can be found in table 2.1.

Table 2.1. Cytogenetic analysis of all samples using SNP-CGH with accompanying FISH diagnosis. Control 1 was selected due to the lack of whole chromosome gains or losses. Autosomes and sex chromosomes are presented separately due to the elevation in LRR described in the discussion. Calculation of "Numeric BAF" is described in the "SNP-CGH" section of the methods. LOH = loss of heterozygosity.

Sample	FISH Diagnosis	SNP Diagnosis	Visual LRR	Average LRR	Visual BAF	Numeric BAF	Reference
control 1	64,XX,t(1;21)	64,XX,t(1;21)	normal X	0.1	normal X	659 A/B SNPs	Lear *et al*. 2008
control 2	63,XO	63,XO	copy loss for X	-0.24	LOH for X	15 A/B SNPs	TLL, unpublished data
control 3	63,XO	63,XO	copy loss for X	-0.21	LOH for X	12 A/B SNPs	TLL, unpublished data
control 4	63,XO	63,XO	copy loss for X	-0.21	LOH for X	11 A/B SNPs	TLL, unpublished data
control 1	64,XX,t(1;21)	64,XX,t(1;21)	normal 31	-0.09	normal 31	0.458 / 0.557	Lear *et al*. 2008
control 5	65,XY,+31	65,XY,+31	copy gain for 31	0.27	split BAF for 31	0.392 / 0.636	Lear *et al*. 1999
control 6	65,XX,+31	65,XX,+31	copy gain for 31	0.13	split BAF for 31	0.357 / 0.607	TLL, unpublished data
case 1	65,XX,+27	65,XX,+27	normal 27	0.08	split BAF for 27	0.432 / 0.566	Figure 2.2
case 2	n/a	65,XY,+30	copy gain for 30	0.23	split BAF for 30	0.337 / 0.639	Figure 2.3

Cytogenetic analysis of the two cases

Chromosomal G-banding from blood lymphocytes in case one showed diploid numbers of 64 and 65, both with two normal X chromosomes. The 2n=65 metaphase spreads accounted for about 20% of the samples and included an additional unidentifiable small acrocentric chromosome. Initial visual examination of the genome-wide SNP-CGH plots showed no alteration of LRR and a slight split in BAF for chromosome 27 (Figure 2.2). The average LRR was 0.079 and BAF averages were 0.432 and 0.566. These values indicated a mosaicism with around 30% of the skin cells carrying an abnormal trisomic karyotype (as discussed in Conlin *et al.* 2010), though due to the noise in the data and the novelty of this technique in the horse, additional confirmation was sought for a diagnosis. In the FISH analysis of the ovary and second skin sample, only the abnormal karyotype of 65,XX,+27 was observed (Figure 2.2).

For case two, visual examination of the genome-wide plots indicated a one copy gain for chromosome 30 (Figure 2.3). Calculated averages were 0.232 for LRR and 0.337 and 0.639 for BAF, corresponding to a 95% mosaicism. No blood was available for karyotype or FISH analysis.

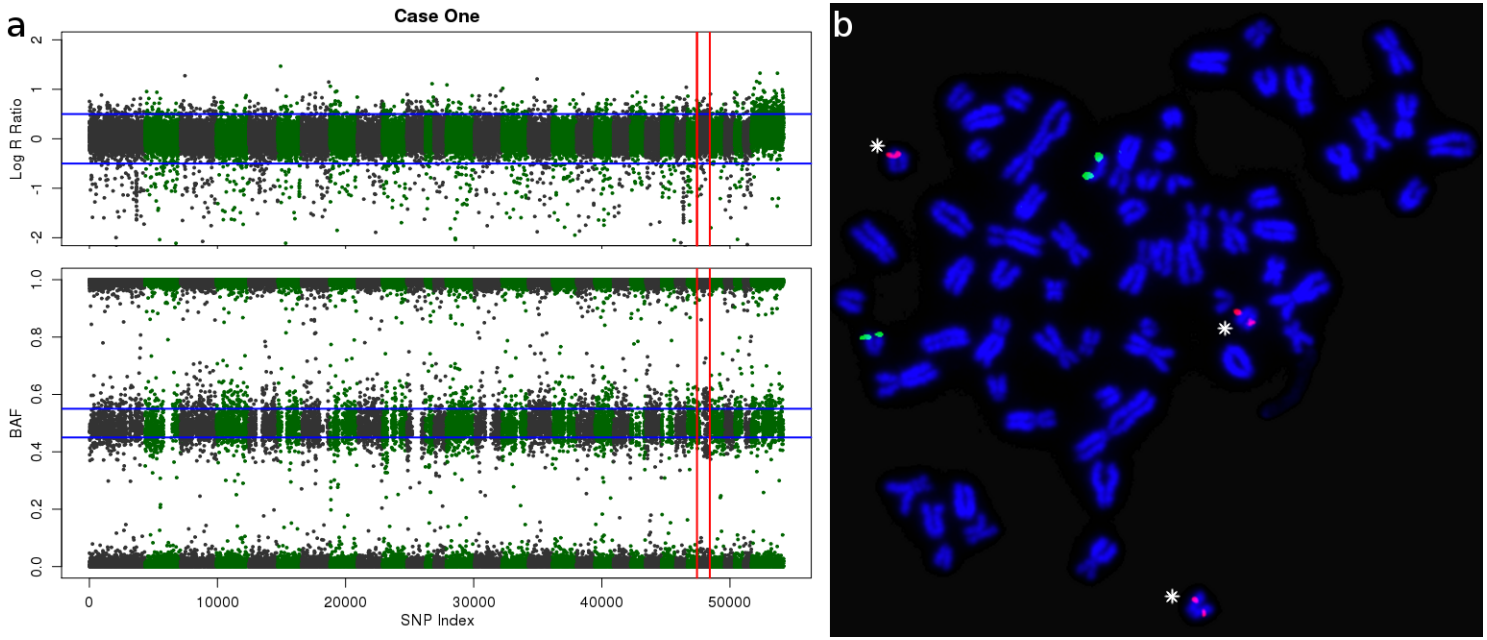

Figure 2.2. Cytogenetic analysis of case one. **a** Genome-wide SNP-CGH plots with chromosomes shown in alternating colors. The vertical lines indicate the location of chromosome 27. While there is no apparent shift in LRR when compared to the other autosomes, the slight split in heterozygous BAF indicates a fraction of cell lines with genotypes AAB and ABB. The increase for LRR on chromosome X is expected as cluster files were generated without separating sexes. **b** FISH screen showing chromosome 30 in green and chromosome 27 in orange (indicated by stars).

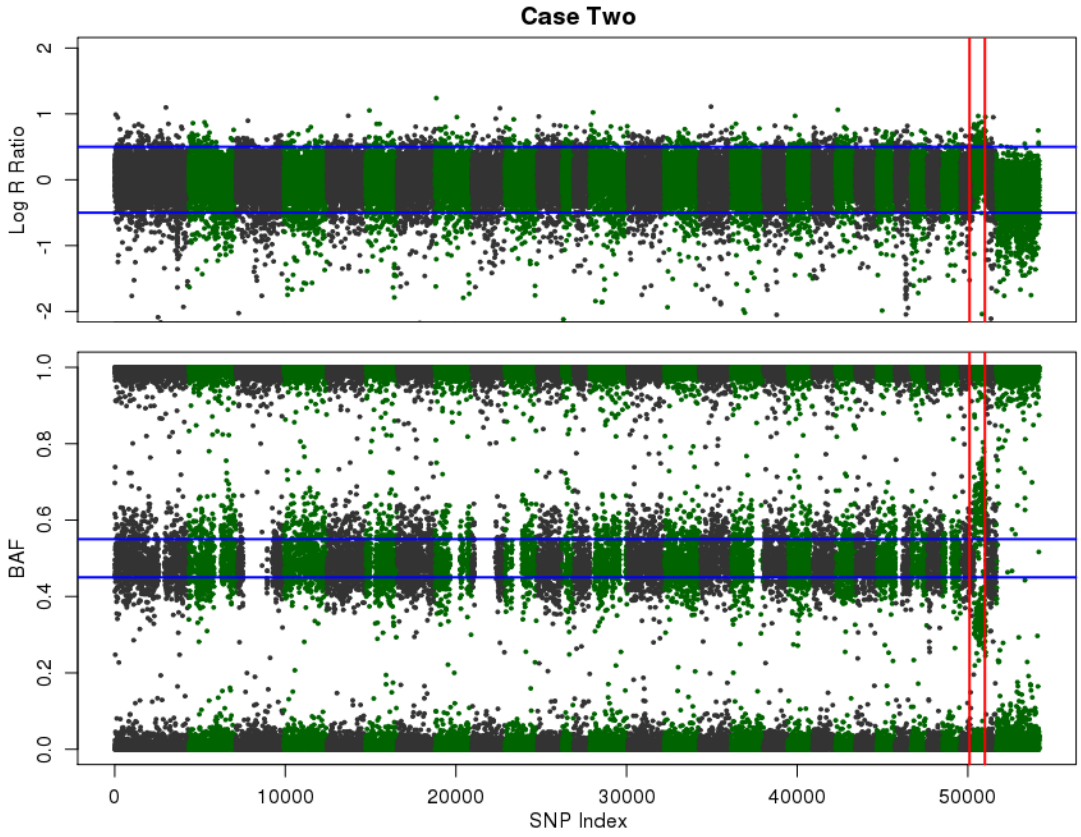

Figure 2.3. Cytogenetic analysis of case two. Chromosome 30 is bordered by vertical lines. The upwards shift in LRR and split in central BAF is indicative of a single copy gain.

DISCUSSION

We successfully used SNP-CGH to identify several cases of aneuploidy in the horse. Both chromosomal copy losses and gains were immediately apparent with visualization of the data. The resulting diagnoses were in agreement with all available FISH data.

In general, all samples appear to show noticeable runs of homozygosity. However, as one mixed breed sample (data not shown) appeared to have shorter and less frequent homozygosity, it is likely an effect of breed structure and low marker density. The Equine SNP50 chip was based largely on SNPs identified in an inbred individual of the Thoroughbred breed. Polymorphisms in this breed may not be representative of other breeds, thus increasing apparent homozygosity. Also, with an average marker spacing of 43.1k, much variation in the genome is not likely to be sampled.

The elevation in LRR seen in the females is likely due to the original generation of the cluster files. For most reference cluster files generated by Illumina, both males and females are used to generate values for all chromosomes, including X (Wang *et al.* 2007). However, as BAF is not affected by the addition of males in the cluster file, it should be weighed more heavily in detected abberations on the X chromosome.

The mosaicism observed in case one was highlighted by regional sampling for DNA. The first skin biopsy was selected for analysis due to the obvious differences in hair color on the filly. However, in the laboratory, individual regions of varying shades (and therefore perhaps varying in karyotype) could not be discerned confidently enough to dissect them individually. Given the later karyotyping showing only abnormal cell lines in the ovary and second skin sample, the SNP-CGH sample likely

included both hair colors with the normal karyotype compromising the majority of the sample.

In case one, the brindle type pattern characterized by striping of two hair coat colors could be due to the presence of a beta-defensin gene cluster on chromosome 27. In dogs, a genetic variant in CBD103 was shown to cause dominant black coat color through binding to the melanocortin-1-receptor (Candille *et al.* 2007). The researchers also created transgenic mice with the normal and variant alleles and found both were capable of inducing pigment-type switching to black in normally agouti-banded hairs (Candille *et al.* 2007). It is thus possible that the extra copy of chromosome 27 (containing the defensin genes) resulted in a higher concentration of melanin in the trisomic section of the filly's hair coat.

Case two likely represents a pure trisomy, despite the LRR and BAF corresponding to a 95% mosaic in the table provided by Conlin *et al.* 2010. In their manuscript, a formula was used to generate this table of expected values at varying levels of mosaicism. However, noise present in genotyping data can result in observed values not matching exactly to expected values (as seen by non-zero LRR in samples with normal karyotypes). As a result, data should be examined for deviations within the sample, since each chromosome should have the same amount of noise.

SNP-CGH represents a viable tool for screening horses with suspected chromosomal abnormalities. While it cannot identify balanced translocations like karyotyping, it offers higher-resolution screens and the ability to detect sub-chromosomal changes in copy number (Srebniak *et al.* 2012). However, resolution is limited by chip marker depth, which does pose problems for identification of short copy number variants in the horse. It is also possible to obtain the necessary DNA for genotyping from lysates of pulled hair roots (Brooks *et al.* 2010). This allows for easy

shipping of samples to the laboratory doing the analysis - hair can be shipped in a normal mailing envelope without need for cold conditions.

ACKNOWLEDGMENTS

The authors thank the owners of the horses in this study for providing samples and Julie Fronczek, Suellen Charter, and the rest of the Cytogenetics Laboratory at the San Diego Zoo's Institute for Conservation Research for cell culture assistance. TLL would like to acknowledge Judy Lundquist for technical assistance.

REFERENCES

Brooks SA, Gabreski N, Miller D, Brisbin A, Brown HE, Streeter C, Mezey J, Cook D, Antczak DF (2010) Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for Lavender Foal Syndrome. PLoS Genet 6: e1000909.

Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, Nix MA, Kerns JA, Schmutz SM, Millhauser GL, Barsh GS (2007) A -defensin mutation causes black coat color in domestic dogs. Science 318: 1418-1423.

Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB (2010) Mechanisms of mosaicism, chimerism, and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Gen 19: 1263-1275.

Faas BH, Feenstra I, Eggink AJ, Kooper AJ, Pfundt R, van Vugt JM, de Leeuw N (2012) Non-targeted whole genome 250K SNP array analysis as replacement for karyotyping in fetuses with structural ultrasound anomalies: evaluation of a one-year experience. Prenat Diagn 32:362-370.

ISCNH (1997) International system for cytogenetic nomenclature of the domestic horse. Chromosome Res 5:433-443.

Lear TL, Brandon R, Bell K (1999) Localization of ten horse microsatellite markers by FISH. Anim Genet 30: 235.

Lear TL, Cox JH, Kennedy GA (1999) Autosomal Trisomy in a Thoroughbred Colt: 65,XY,+31. Equine Vet J 31: 85-88.

Lear TL, Brandon R, Piumi F, Terry RR, Guerin G, Thomas S, Bailey E (2001) Mapping of 31 horse genes in BACs by FISH. Chromosome Res 9: 261-262.

Lear TL, Bailey E (2008) Equine clinical cytogenetics: the past and future. Cytogenet Genome Res 120: 42-49.

Lear TL, Lundquist J, Zent WW, Fishback WD Jr, Clark A (2008) Three autosomal chromosome translocations associated with repeated early embryonic loss

(REEL) in the domestic horse (Equus caballus). Cytogenet Genome Res 120:117-122.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Distl O, Guérin G, Hasegawa T, Hill EW, Leeb T, Lindgren G, Penedo MC, Røed KH, Ryder OA, Swinburne JE, Tozaki T, Valberg SJ, Vaudin M, Lindblad-Toh K, Wade CM, Mickelson JR (2012) A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. PLoS Genet 8: e1002451.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-1148.

R Development Core Team (2009) ISBN 3-900051-07-0, http://www.R-project.org/

Seabright M (1971) A rapid banding technique for human chromosomes. Lancet 2:971-972.

Srebniak MI, Boter M, Oudesluijs GO, Cohen-Overbeek T, Govaerts LC, Diderich KE, Oegema R, Knapen MF, van de Laar IM, Joosten M, Van Opstal D, Galjaard RH (2012) Genomic SNP array as a gold standard for prenatal diagnosis of foetal ultrasound abnormalities. Mol Cytogenet 5:14.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665-1674.

APPENDIX

UNPUBLISHED DATA

METHODS AND MATERIALS

Sample collection and case histories

Case three was a 14 year old Thoroughbred mare with a history of reproductive issues. She produced one normal female offspring by one sire and three male offspring with choanal atresia by a different sire. Blood from the mare and hair from one male foal were submitted for karyotyping and cytogenetic analysis. DNA extraction from blood was completed as described above. Hair lysis was performed as first described by Locke *et al*. 2003.

RESULTS

Cytogenetic analysis

Standard karyotyping failed to identify any chromosomal abnormalities in the mare. However, visual inspection of SNP-CGH plots identified a cluster of abnormal SNPs on the X chromosome. This cluster was comprised 27 SNPs spanning 1 Mbp with an average LRR of -0.307 (with the rest of ECAX having an average LRR of 0.198) and no heterozygous BAFs, indicative of a single copy loss (Figure 2.4). There were 4 genes located within this region (*TBX22*, *CHMP1B*, *FAM46D*, *BRWD3*).

Blood was not available from the offspring for karyotyping. SNP-CGH plots (derived from hair DNA) had a high degree of noise, however the same cluster of SNPs appeared abnormal in the colt. The average LRR was -3.51 (compared to -1.15 for all other ECAX SNPs), though BAFs ranged from 0.00 to 0.66.

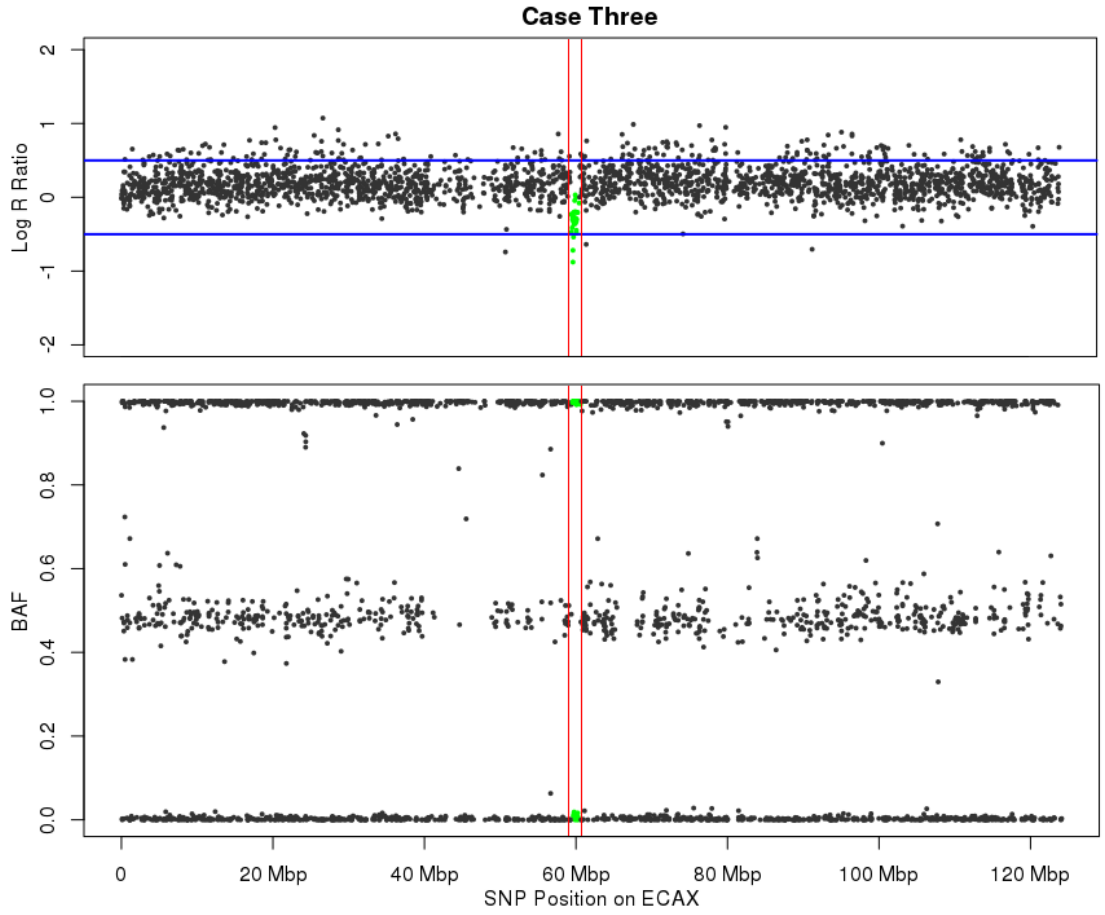

Figure 2.4. Cytogenetic analysis of chromosome X in case three. The vertical lines indicate the location of the 27 SNPs involved in the putative copy number alteration. The lower shift of LRR and lack of heterozygous SNPs compared to the rest of ECAX are indicative of a single copy deletion.

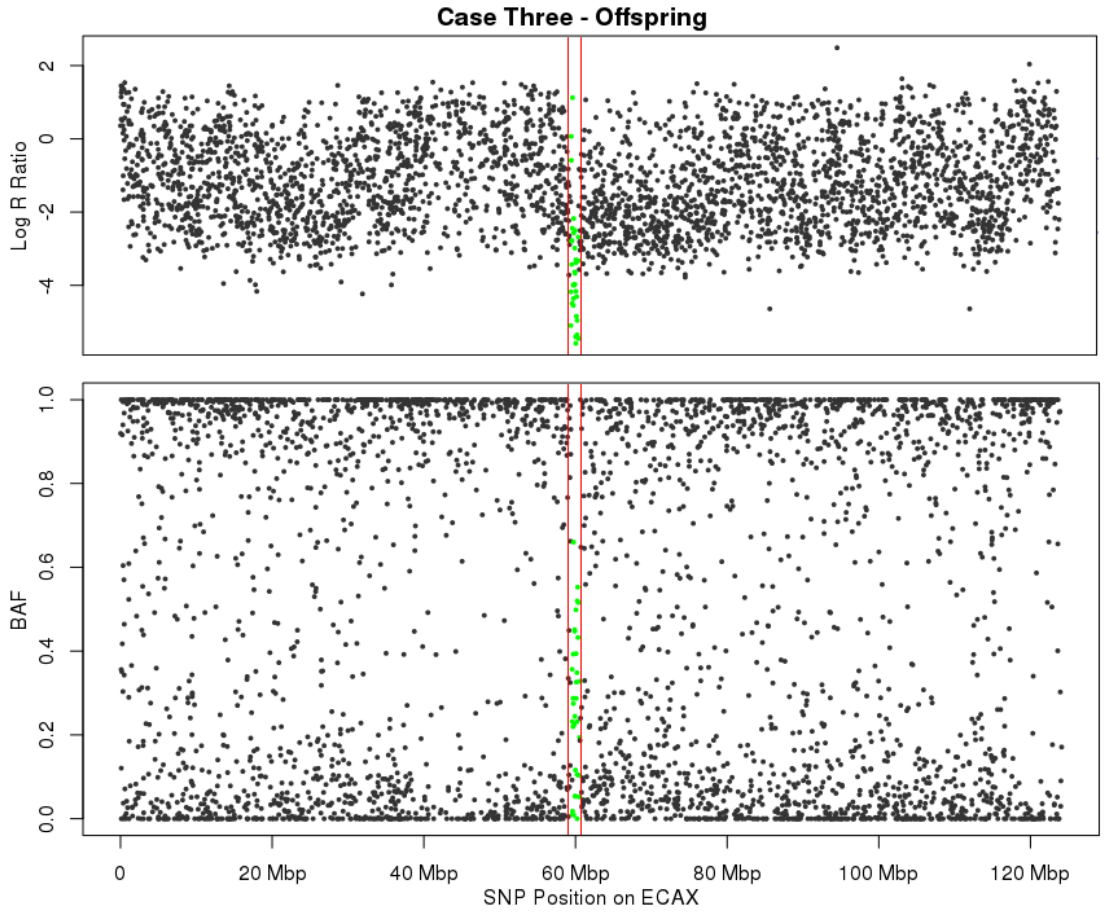

Figure 2.5. Cytogenetic analysis of chromosome X in the offspring of case three. The vertical lines highlight the same region as shown in figure 2.4. The lower shift of LRR compared to the rest of ECAX are indicative of a copy number loss.

DISCUSSION

Choanal atresia is a congenital abnormality in which there is a blockage between the nasal passages and pharynx, resulting in a unilateral or bilateral disruption of airflow (James *et al*. 2006). Studies in other species have identified a small number of genes associated with the phenotype (Reed *et al*. 2010).

Although there are only four offspring in this kindred, the distribution of the disease suggested an issue with the X chromosome. The results seen through SNP-CGH analysis indicated a single copy deletion that was too small to be resolved through conventional karyotyping. The X chromosome bearing this deletion could have been passed on to all three male offspring, resulting in a haploinsufficiency responsible for the abnormality. The reduced logR ratios observed in the male offspring in the same cluster of SNPs support this theory. The high degree of noise in this individual's CGH plots may be due to lower quality DNA from hair. The moderate BAF values is likely an artifact of numeric transformation on intensity values that likely would have been removed from standard filtering by GenomeStudio.

Of the four genes annotated in the deletion region, two have no known function. The third, *bromodomain and WD repeat domain containing 3* (*BRWD3*), is associated with macrocephaly and intellectual disabilities in humans (Field *et al*. 2007). The fourth, *T-box 22* (*TBX22*), is associated with X-linked cleft palate and ankyloglossia in humans, as well as choanal atresia in mice (Kantaputra *et al*. 2011, Pauws *et al*. 2009). Given these phenotypes in other species, and the distribution of phenotypes in case three's offspring, it is highly likely the observed deletion is responsible for the mare's reproductive issues and deaths of the three still-born colts.

This study demonstrates the power of using array CGH for detecting smaller chromosomal aberrations. Conventional karyotyping techniques are limited to alterations visible on the metaphase chromosomes, or to screening using specific FISH

probes. Here, we show even with the low density of the Equine SNP50 chip, we are able to detect a 1 Mbp deletion, one well below the range of detection with conventional methods.

REFERENCES

Field M, Tarpey PS, Smith R, Edkins S, O'Meara S, Stevens C, Tofts C, Teague J, Butler A, Dicks E, Barthorpe S, Buck G, Cole J, Gray K, Halliday K, Hills K, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Varian J, West S, Widaa S, Mallya U, Wooster R, Moon J, Luo Y, Hughes H, Shaw M, Friend KL, Corbett M, Turner G, Partington M, Mulley J, Bobrow M, Schwartz C, Stevenson R, Gecz J, Stratton MR, Futreal PA, Raymond FL (2007) Mutations in the BRWD3 gene cause X-linked mental retardation associated with macrocephaly. Am J Hum Genet 81: 367-374.

James FM, Parente EJ, Palmer JE (2006) Management of bilateral choanal atresia in a foal. J Am Vet Med Assoc 229: 1784-1789.

Kantaputra PN, Paramee M, Kaewkhampa A, Hoshino A, Lees M, McEntagart M, Masrour N, Moore GE, Pauws E, Stanier P (2011) Cleft lip with cleft palate, ankyloglossia, and hypodontia are associated with TBX22 mutations. J Dent Res 90: 450-455.

Locke MM, Penedo MC, Bricker SJ, Millon LV, Murray JD (2002) Linkage of the grey coat colour locus to microsatellites on horse chromosome 25. Anim Genet 33: 329–337.

Pauws E, Hoshino A, Bentley L, Prajapati S, Keller C, Hammond P, Martinez-Barbera JP, Moore GE, Stanier P (2009) Tbx22null mice have a submucous cleft palate due to reduced palatal bone formation and also display ankyloglossia and choanal atresia phenotypes. Hum Mol Genet 18: 4171-4179.

Reed KM, Bauer MM, Mendoza KM, Armién AG (2010) A candidate gene for choanal atresia in alpaca. Genome 53: 224-230.

CHAPTER 3

GENERATION OF A *DE NOVO* TRANSCRIPTOME FROM EQUINE LAMELLAR TISSUE

H. M. Holl[1], S. Gao[2], Z. Fei[2], C. Salter[1], and S. A. Brooks[3*]

[1]Department of Animal Science, Cornell University, Ithaca, NY 14853, USA

[2]Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853, USA

[3]Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

[*]Corresponding author

ABSTRACT

The equine hoof is a specialized structure in which the distal skeleton is suspended within the capsule by interdigitated structures known as laminae. Inflammation of this tissue, known as laminitis, is a devastating disease that is the second leading cause of both lameness and euthanasia in the horse. Current research on the laminitic transcriptome focuses on the expression of known genes. However, as this tissue is quite unique and equine annotation is largely derived from computational predictions and gene models from other species, there are likely yet uncharacterized transcripts expressed in the laminae that may be involved in the etiology of laminitis. In order to create a novel annotation resource, we performed whole transcriptome sequencing of sagittal lamellar sections from one control and two laminitis affected horses. Assembly of 113 million 100bp reads resulted in around 75,000 transcripts. Of these, 36,000 corresponded to known annotation in NCBI's non-redundant protein database. RT-PCR of 12 selected assemblies confirmed structure and expression in lamellar tissue. Transcriptome sequencing represents a powerful tool to expand on equine annotation and identify novel targets for further laminitis research.

INTRODUCTION

Laminae are interdigitated dermal and epidermal tissues found in the hooves of livestock that form the attachment to the distal skeleton. Equids have an additional specialization in the form of secondary laminae that project from the primary laminae which further increase the surface area and thus strengthen this connection (Pollitt 2010). The junction between dermal and epidermal laminae must be strong enough to withstand the forces of weight bearing and motion without separation, while providing sufficient flexibility to absorb concussive forces and allow growth. Inflammation of the laminae (laminitis) is a devastating disease that can lead to separation of these tissues and a rotation of the third phalanx (P3) away from the hoof wall.

The etiology of laminitis is poorly understood. Many risk factors have been identified in the horse, including inflammation in other parts of the body, sepsis, metabolic conditions, or mechanical stress (Eades 2010). Currently, as there are very few treatments available, prevention through avoiding known risk factors is recommended. In the early stages of laminitis (either pre-clinical symptoms or at the onset of lameness), prolonged cooling of the hooves in ice water has been shown to reduce severity of the disease and prevent separation of the laminae (van Eps *et al*. 2013). However, if adequate treatment is not provided promptly, euthanasia is often the result. A study from the USDA in 1998 estimated the annual cost of lameness at $678 million, with laminitis accounting for 15% of the reported cases (NAHMS 2001). The American Association of Equine Practitioners has specifically identified laminitis as the disease most frequently reported as needing more research.

Several methods have been devised to experimentally induce laminitis, including carbohydrate overload, oligofructose overload, and black walnut extract administration. Although all of these models will result in the disease, key differences in physiological response (as compared to the natural etiology) have been

demonstrated (Faleiros *et al*. 2011 , Leise *et al*. 2011). However, as natural cases can be much more difficult to acquire, these models continue to serve an important role in research.

Gene expression has been applied in studies to better understand the disease process. However, much of this research has focused on the expression of few known genes, using qPCR to target specific pathways (Kwon *et al*. 2013, Steelman *et al*. 2013, Wang *et al*. 2013, Wang *et al*. 2014). Only two studies have attempted a transcriptome-wide view of laminitis. The first commercially available whole-transcriptome equine-specific microarray was not published until 2009, therefore early studies attempted two different approaches. The first study chose to use cross-species hybridization with the bovine gene expression chip, identifying 155 out of the 15,000 genes assayed to be significantly up-regulated (Budak *et al*. 2009). They were unable to identify any down-regulated genes, which was likely due to the high false-negative rate associated with imperfect hybridization. A second study instead generated a custom equine-specific array with 3076 targets derived from leukocyte EST libraries (Noschka *et al*. 2009). Less than 100 of these genes were found to have significant differential expression.

Both of these projects, and any current work utilizing microarrays, are hindered by insufficient genome annotation in the horse. The only major annotation attempt used an older sequencing technology, generating 35 bp reads from eight diverse tissue types (Coleman *et al*. 2010). They identified that 48% of genes displayed tissue-specific expression patterns, with 7% of the genes only found in one tissue type. However, this data was not incorporated into automatic annotation pipelines for the popular genome browsers, and lamellar tissue was not included in sequencing. Using this data, the authors also demonstrated there were 428 genes

completely lacking in equine annotation, even though many of these genes have data in other species (Coleman *et al*. 2013).

Whole transcriptome sequencing (RNA-seq) is a promising solution for interrogation of gene structure and expression, especially in a divergent tissue like the hoof. RNA-seq is a hypothesis-free examination of all cDNA in a given sample, allowing for the identification of unique features such as unannotated transcription, splice sites, allele-specific expression, anti-sense expression, and alternative poly-adenylation (t' Hoen *et al*. 2008, Malone and Oliver 2011, Wilson *et al*. 2014). Additionally, technical variation is reportedly low, with high reproducibility between lanes (Marioni *et al*. 2008). Studies have continuously demonstrated high correlation between microarray differential expression studies and RNA-seq strategies, noting the main difference is improved sensitivity for low-abundance transcripts by RNA-seq (Mooney *et al*. 2013, Zhao *et al*. 2014). However, as RNA-seq is still considerably more expensive and computationally intense than microarrays, much mainstream research still relies on microarrays or qPCR.

The objective of this study was to produce a transcriptome resource for the study of laminitis. Given that recent studies rely heavily on qPCR, the generation of a set of equine, hoof-specific transcripts can greatly benefit in the selection of novel targets for expression studies. Current annotation is largely based on computational predictions and gene models from other species, among which there is not a good physiological model for the laminae. Additionally, while there have been a few equine RNA-seq studies, raw data is often only placed in public databases and not fully processed or curated (Coleman *et al*. 2010, Park *et al*. 2012, Capomaccio *et al*. 2013, Igbal *et al*. 2014). Thus these valuable datasets are difficult to access and may require intensive bioinformatic analysis before use in subsequent projects, and sadly are often underutilized.

MATERIALS AND METHODS

Sample collection and transcriptome sequencing

Samples were collected from horses sent to the necropsy unit at the Cornell University Veterinary School. Medical history was collected whenever available. Hoof sections were placed on ice for transport to the lab, where a histological examination was performed and mid-sagittal lamellar sections were dissected out. Samples were placed into RNA later and stored at -80°C until ready for processing.

Phenotype was assessed through medical history, physical exam prior to euthanasia, and histological findings. Control animals were defined by P3 running parallel to the hoof wall with no bruising or thickening of the laminae. Acute cases often had some degree of rotation, as well as profuse bleeding and thickened laminae. Chronic cases were defined by thickened, fibrous lamina with bruising throughout, and sometimes with pooled blood by P3. Sample information can be found in Table 3.1.

RNA was extracted from approximately 60 mg of lamellar tissue using the Qiagen RNeasy kit (Qiagen Inc., Valencia, CA, USA) following manufacturer's protocols for fibrous tissue. 50 µL of RNA was DNase treated using either the Ambion Turbo DNA free kit (Life Technologies, Carlsbad, CA, USA) or Qiagen DNase I kit, followed by Qiagen RNA cleanup kit. Quantification was carried out using a NanoDrop spectrophotometer (NanoDrop Technologies LLC., Wilmington, DE, USA).

Library preparation and sequencing was performed by Cornell University's Life Sciences Core Laboratory Center. Single-end libraries were constructed using manufacturer's protocols for poly-T selection and sequenced on an Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA). Raw reads were submitted to the European Nucleotide Archive (accession number PRJEB6100).

Table 3.1. Summary of samples used in this study. Laminitis phenotype was determined through medical history and histological examination.

Sample	Phenotype	History	Experiment
CU1	control	healthy	RNA-seq, RT-PCR
CU5	control	healthy	RT-PCR
CU17	acute	enterocolitis	RT-PCR
CU18	acute	enterocolitis	RNA-seq, RT-PCR
LSU-E	chronic	Equine Metabolic Syndrome	RT-PCR
LSU-J	chronic	Equine Metabolic Syndrome	RNA-seq, RT-PCR

De novo *assembly*

Raw RNA-seq reads were processed in two steps. First, a custom R script (based on the ShortRead package) was used to remove adapter and barcode sequences, as well as to trim low quality (Q<20) bases from both ends of the reads (Morgan *et al*. 2009). Trimmed reads shorter than 25 bp were discarded. Second, reads were aligned to the GenBank virus (version 186) and ribosomal RNA sequence databases with BWA under default parameters (Li and Durbin 2010). Only unmapped reads were retained for assembly.

The filtered reads from all samples were pooled and *de novo* assembled into contigs using Trinity with "min_kmer_cov" set to 2 (Friedman *et al*. 2011). In order to remove some of the redundancy of Trinity-generated contigs, a further assembly step using iAssembler with a minimum of 99% identity (-p) was performed (Zheng *et al*. 2011). Contigs shorter than 200 bp were discarded.

Unigene annotation

All unique transcripts (unigenes) were compared to the GenBank non-redundant protein database using blastx with an E-value cutoff of 1e-5. Only the protein with the highest E-value (and thus highest significance) was retained for further analysis.

Unigenes were also aligned to the equCab 2.0 reference genome using BLAT with parameters recommended for same-species mRNA alignments (Kent 2002). The pslCDnaFilter tool was used to remove alignments with less than 200 bp, 98% identity, or 50% coverage. The resulting PSL file was converted to BED format and compared with Equine-specific repeat annotation using BEDtools intersectBed in order to filter out alignments that contained over 10% repetitive DNA (Pruitt *et al*. 2014, Quinlan and Hall 2010). Many retroviruses in the genome are expressed, but

high homology among these elements often leads to chimeric and spurious assemblies, and thus creates problems for alignment-based analyses. The filtered unigenes were then compared to NCBI Non-Horse RefSeq, Horse RefSeq, and Horse Ensembl annotations using intersectBed at 10% overlap.

Putative gene names were assigned to unigenes based on high quality matches to NCBI non-redundant databases. Two BED files were produced for use in genome browsers (one containing all transcripts and one with only large transcripts containing 3 or more exons) and are available at http://www.animalgenome.org/repository/horse/.

Variant calling

Raw sequencing reads were aligned to the EquCab 2.0 reference genome using BWA under default parameters. SAMtools was used to convert alignments to BAM format and to remove PCR duplicate reads (Li et al. 2009). SNPs were identified with GATK using the recommended pipelines with a Q>30 cutoff (McKenna et al. 2010, DePristo et al. 2011, Van der Auwera et al. 2013). VCFtools was then used to filter out variants with fewer than 10 observations, followed by BEDtools to remove variants that fell outside of regions with corresponding assembly alignments (Danecek *et al*. 2011). The final list of variants was pooled and submitted to NCBI dbSNP.

Analysis of putative novel loci

We screened the transcriptome assembly for novel loci with two steps. First, a second genome alignment was prepared by running RepeatMasker (using RepBase 2013-04-22 libraries) on the unigenes, then BLAT and subsequent filtering was performed as before (Smit *et al*. 2010, Jurka *et al*. 2005). Next, the unmasked and masked alignments were compared, and unigenes that passed filtering criteria in both datasets were selected. The unmasked alignments of these unigenes were then

compared to RefSeq annotation using BEDtools, and alignments with less than 5% overlap to known annotation were labeled as putative novel loci. All matches to the unassembled chromosome (chrUn) were discarded. Although valuable novel genes are likely to be found there, the incomplete state of assembly in this region makes downstream alignment based analyses problematic.

Twelve novel genes were selected for RT-PCR validation and proof of concept based on additional criteria. ExPasy “translate” tool was used to identify open reading frames (ORFs) in these unigenes (Gasteiger *et al*. 2003). These were then BLATed back to the equCab 2.0 reference genome, and only unigenes with ORFs spanning at least three exons on their corresponding transcript annotation were kept, thus identifying larger transcripts with significant exon/intron structure. The ORFs were then compared to the non-redundant protein database using blastp, and targets with little to no experimental data were selected for further validation.

Within each gene, an amplicon of cDNA was targeted using intron spanning primers created with the Primer3 software (Table 3.2, Rozen and Skaletsky 1998). Two-step RT-PCR was performed in 15 μL reactions using the SuperScript VILO MasterMix kit (LifeTechnologies, Carlsbad, CA, USA) followed by standard PCR. 1 μL of cDNA was amplified in 10 μL PCR with FastStart Taq DNA polymerase (Roche Applied Science, Branford, CT, USA) and included all reagents per the manufacturers recommended conditions.

Amplification was verified on 3% agarose gel, and the resulting PCR products were submitted to the Cornell Core Life Sciences Laboratories Center for sequencing using standard ABI chemistry on a 3730 DNA Analyzer (Applied Biosystems Inc., Foster City, CA, USA). Amplicons were aligned to their corresponding unigenes to confirm identity using Consed (Gordon *et al*. 1998).

Table 3.2. Primers used to confirm expression of unannotated transcripts. All PCRs were performed with an annealing temperature of 62°C and an elongation time of 30s.

Name	Forward Seq	Reverse Seq	Size
UN20159	TTCAAGAGCAATGGGATGCT	CGCAGTGTCATGAACAGGTTA	227 bp
UN14299	TTTTCCTCTGAAGCATTTCC	TAGAGCATCGCTTTCCTGGT	284 bp
UN30143	CCCACCCCCAACCTAGATAC	AGGTAAGACAGGCTGGGTCA	499 bp
UN27297	GTCCGAATTCAGCCAATCAT	GAAACGATTTATGGCCTCCA	495 bp
UN27113	TGAAAGGCATCCATCTGGTC	ACCCCGTTACAGAGGTCCTT	329 bp
UN28086	TCCTTGCTAGGATGCTCTGG	GAGCACCAGGATGAAGAGGA	506 bp
UN62514	GGCTCCTCCTCCTTGTGAG	AACAGCAGTTTGGCAGGAGT	437 bp
UN21936	CTATGTTCTGGGCTGTGGTG	TGTAGCCACGTTTGCACTCT	485 bp
UN70945	CCTCATGACCTTCGTGGTTC	ATCTTTTTGAGCTGGCAAGG	409 bp
UN26965	GCACCCTACTCCCACATACG	GCTCACATCCACGTCTGCTA	422 bp
UN26584	GTACATTCCTCCCCTGCAAA	TCGACACCATCCAGTTGAAA	479 bp
UN50658	CTGACCAGGACCCTCAGTCT	TCAGTGACCAGGCCTTCTTC	343 bp

RESULTS

Illumina sequencing and assembly

Whole transcriptome sequencing of the three samples in this experiment generated in a total of 112,979,003 reads. Sequencing data from all three individuals was pooled for assembly in order to capture genes that may be rare or unique to the laminitic state. After filtering, 87,598,529 high-quality reads remained. A summary of assembly metrics can be found in Table 3.3. The number of unigenes mapped per locus ranged from 1 to 139, averaging 2.44 isoforms representing 25,580 loci. Many of these unigenes are shorter transcripts covering only a single exon or splice junction, partially due to lowly expression transcripts lacking sufficient coverage for assembly (Figure 3.1). Considering only the longer 3+ exon transcripts resulted similar statistics (Table 3.4).

Overall, 88% of raw sequencing reads mapped to the genome. The GATK recommended pipeline identified a total of 131,034 SNPs. In order to account for false positives resulting from incorrectly mapped spliced reads, the transcriptome assembly was used to filter out SNP calls that fell outside of the repeat-filtered annotation generated by this assembly. The 119,555 SNPs that remained (91.2%) were submitted to public databases (Table 3.5).

Table 3.3. D*e novo* assembly statistics.

Metrics	Raw Assembly
Total reads (100 bp)	112,979,003
Reads after filtering	86,275,849
Average read length after filtering	88.3 bp
# Unigenes	74,860
N50	2,272
Minimum Length	201
Average Length	1,098
Maximum Length	17,667

Table 3.4. Isoform statistics by locus. Unigenes are clustered together based on an overlap of at least 1 bp.

Statistics	All Transcripts	Long (3+ Exon) Transcripts
Total Unigenes	55,120	27,884
Unique Loci	23,779	12,905
Min Unigenes per Locus	1	1
Max Unigenes per Locus	125	89
Average Unigenes per Locus	2.32	2.16

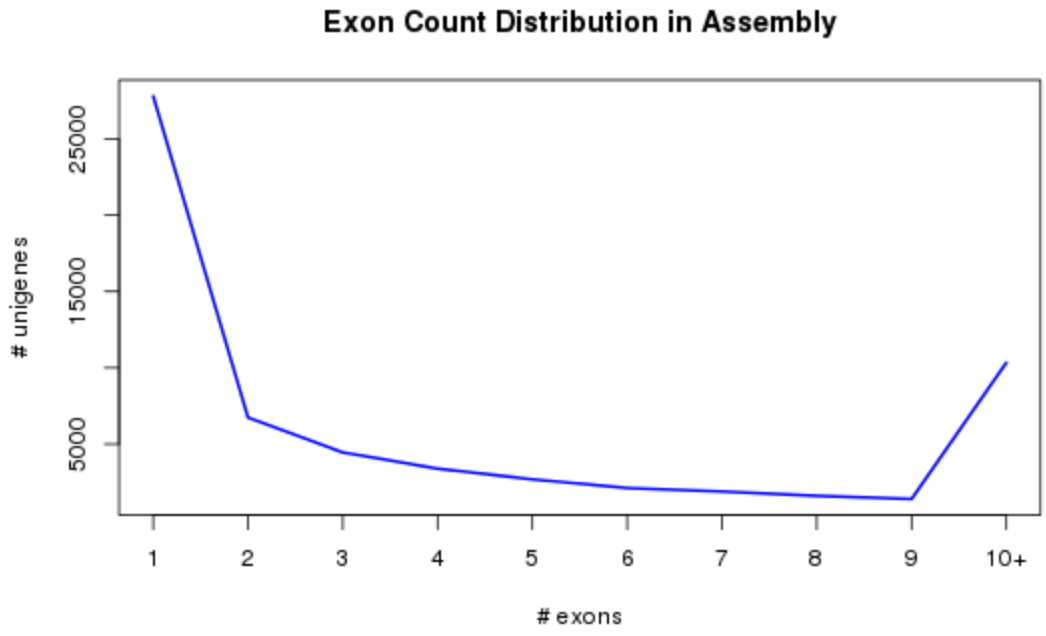

Figure 3.1. Distribution of exon counts within the unfiltered assembly. Longer models range from 10 to 119 exons.

Table 3.5. Mapping statistics for RNA-seq onto EquCab2.

Sample	Phenotype	Total Reads	Mapped Reads	% Mapped	SNPs
CU1	control	36,277,643	31,561,549	87%	60,652
CU18	acute	43,422,463	38,211,767	88%	72,364
LSU-J	chronic	33,278,897	29,618,218	89%	58,414

Annotation with known gene and protein databases

Using blastx, a total of 36,195 unigenes (48%) had significant matches to proteins in the non-redundant database. To simplify the analysis, only the top hits were retained. 35% of the matches were to equine proteins, and of these, 97% were computationally derived entries (XP_ accession numbers).

Additionally, unigenes aligned by BLAT to the equine genome were compared to the NCBI horse RefSeq, NCBI non-horse RefSeq, and Ensembl prediction tracks available from the UCSC Genome Browser. A summary of overlap between the known databases is provided in Table 3.6.

Gene IDs were assigned to each unigene based on matches to the non-redundant protein database or RefSeq alignments, resulting in annotation of 44,730 transcripts. Unannotated transcripts retained their identifier provided by Trinity. These transcripts likely correspond to novel genes or non-coding RNA and were selected for further examination. An example of the generated custom track annotation is found in Figure 3.2.

Table 3.6. Unigenes matching annotation in various databases. The repeat-filtered assembly was utilized for EquCab2 alignment-based annotation.

Database	Total Records	Unigenes
NCBI NR Protein	37,818,139	36,195 / 74,860[a] (48%)
Equine-Specific Repeats	2,905,169	19,740 / 74,860[a] (26%)
Non-Horse RefSeq	255,606	24,501 / 55,120[b] (44%)
Ensembl Predictions	29,196	15,538 / 55,120[b] (28%)
Horse RefSeq	1,169	604 / 55,120[b] (1%)
None	n/a	31,091 / 74,860[a] (42%)

[a]Unfiltered transcriptome assembly

[b]Repeat-filtered transcriptome assembly

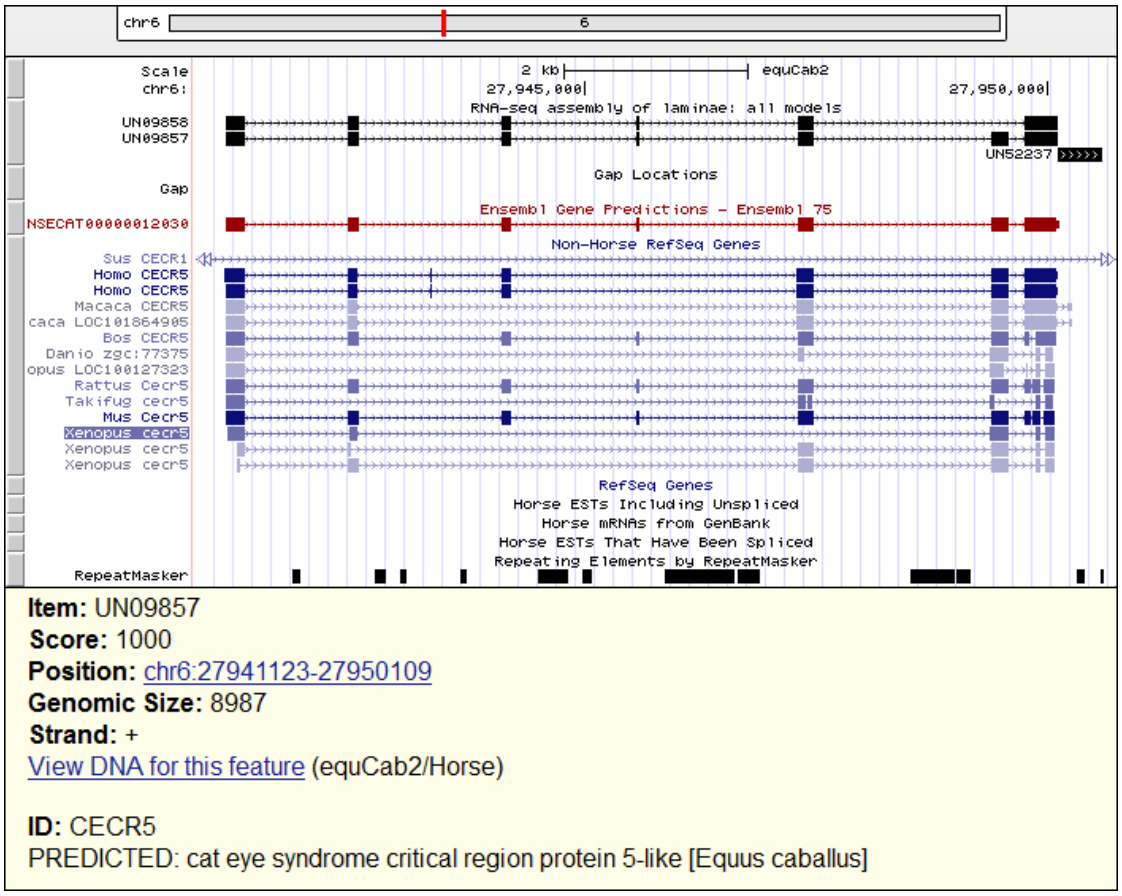

Figure 3.2. Example custom annotation on UCSC Genome Browser. Clicking on the identifier “UN09857” loads the table in the lower panel. Custom identifiers provide corresponding gene and protein annotation for each unigene.

Amplification and sequencing of cDNA from putative transcripts

There were a total of 13,632 unigenes with 3 or more exons that did not match to known annotation. Of these, there were 4,718 that did not overlap with other unigenes. A subset of 12 unique transcripts that contained ORFs which spanned over 3 exons were selected for molecular validation (Table 3.7).

RT-PCR successfully amplified cDNA from all selected transcripts. All products were of the expected length and Sanger-derived sequences matched completely with assembled sequences. As differential expression was not the goal of this study, no quantitative analyses were attempted. However, one selected transcript did display a qualitative trend for disease-specific expression (Figure 3.3). The best match (placenta-specific protein 1 precursor), located on ECAX, is a computational prediction with support from 1 equine mRNA and 85% coverage of RNA-seq alignments from one sample in the short-read archive. The only other equine match was to a homologous gene, placenta-specific protein 1-like (E=4e-9), which was mapped approximately 100 kb downstream of the alignment on chromosome 12. However, this record is completely computationally derived, supported only by similarity to two proteins.

Table 3.7. Putative novel loci examined by RT-PCR. Gene models of two unigenes can be found in Figures 3.4 and 3.5.

Name	Chr	Start	End	ORF	Ex	E	BLASTX
UN20159	1	160767946	160771835	85	4	1.2E-40	AAA80518 T-cell receptor alpha chain (IgC TCRA) [Equus caballus]
UN14299	4	24422341	24427035	83	4	none	none
UN30143	5	15325547	15363449	515	16	7.3E-22	XP_003209297 PREDICTED: intraflagellar transport protein 80 homolog [Meleagris gallopavo]
UN27297	5	74952780	74965671	177	6	1.7E-56	XP_001493637 PREDICTED: guanylate-binding protein 5 [Equus caballus]
UN27113	9	81709438	81711119	137	4	2.9E-60	XP_001917082 PREDICTED: lymphocyte antigen 6H-like [Equus caballus]
UN28086	11	880333	894532	240	6	9.9E-31	NP_663348 secreted and transmembrane protein 1A precursor [Mus musculus]
UN62514	11	37045433	37058857	162	5	1.5E-11	ACI67873163 Perlwapin [Salmo salar]
UN21936	12	19828520	19834360	195	5	2.1E-14	NP_001243909 placenta-specific protein 1 precursor [Equus caballus]
UN70945	22	34309639	34312469	141	4	1.7E-54	AES10462 antileukoproteinase-like protein [Mustela putorius furo]
UN26965	24	44802220	44806274	142	5	3.5E-21	XP_002696828 PREDICTED: uncharacterized protein LOC509029 [Bos taurus]
UN26584	28	21567866	21620053	248	7	5.8E-102	XP_003952339 PREDICTED: uncharacterized protein LOC101059192 [Pan troglodytes]
UN50658	X	2231083	2240297	174	7	1.0E-38	XP_003134963 PREDICTED: odorant-binding protein-like [Sus scrofa]

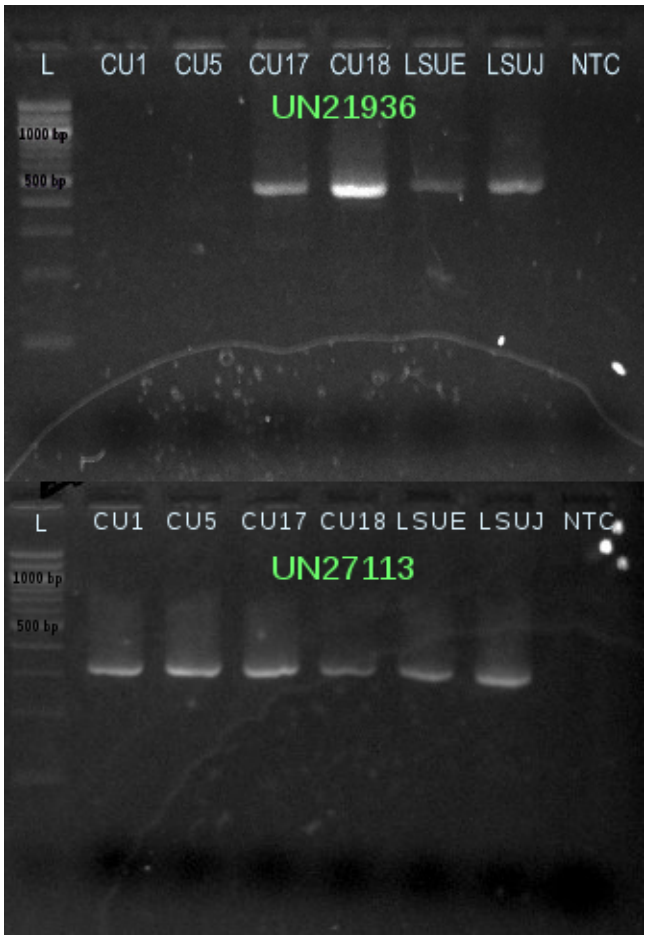

Figure 3.3. Agarose gel demonstrating the expression of UN21936 and UN27113. Expression of UN21936 appears to be limited to laminitic samples. CU1, CU5 = control; CU17, CU18 = acute laminitis; LSUE, LSUJ = chronic laminitis; NTC = non-template (negative) control.

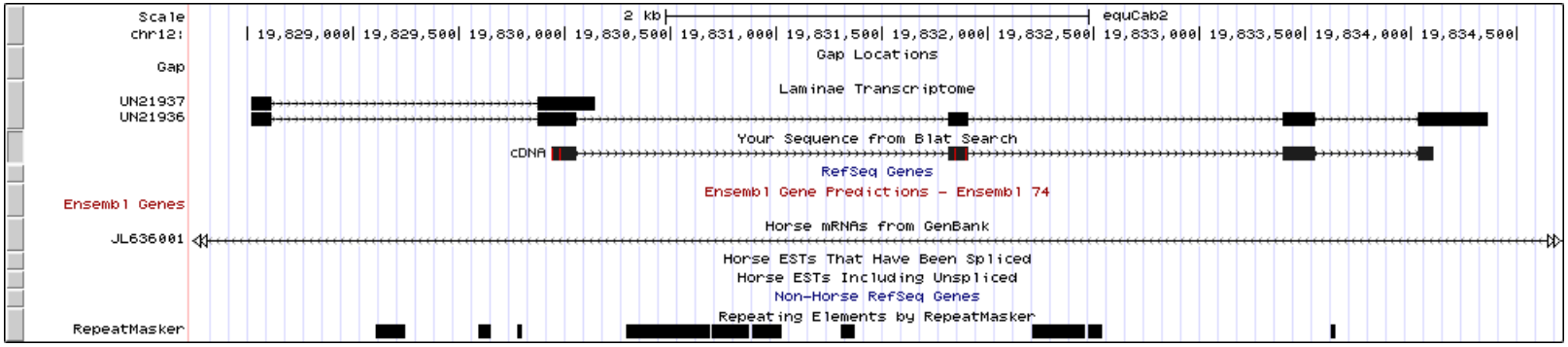

Figure 3.4. Alignment of sequenced cDNA from UN21936 and assembled transcripts to the reference genome. Screenshot was captured from the UCSC Genome Browser. Dark boxes represent exons while thin lines are introns. The empty RefSeq Genes, Ensembl Gene Predictions, Horse ESTs, and Non-Horse RefSeq tracks indicate that there has never been expression or computational predictions placed here. Although the amplicon shows three mismatches to the reference, this sequence aligned perfectly to the assembled transcript.

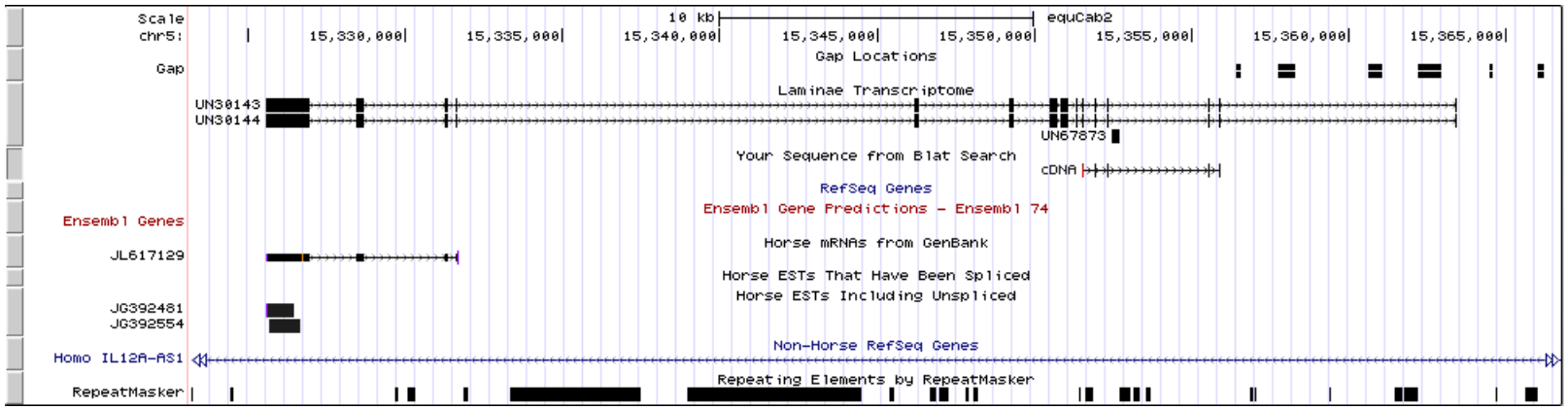

Figure 3.5. Alignment of sequenced cDNA from UN30143 and assembled transcripts to the reference genome. Screenshot was captured from the UCSC Genome Browser. One mRNA and two cDNAs in public databases match part of the aligned transcript, though no gene model exists. This transcript was the longest model of the unannotated list.

DISCUSSION

We utilized RNA-seq to successfully generate a transcriptome assembly of equine lamellar tissue. As the hoof is a specialized tissue, it likely has unique transcripts that previous annotation efforts would have missed. This data set represents a valuable tool for laminitis research, providing information on both known genes expressed in the hoof, as well as a wealth of previously unannotated transcripts. The transcripts identified in this study can now be utilized with other technologies to search for novel targets with relevance to laminitis.

RNA-seq provides unprecedented power for transcript and isoform discovery. However, relatively little of this information trickles down in to human readable annotation and applied datasets useful to the average molecular biologist. While some resources now exist that attempt to bridge this gap by providing bioinformatics instruction for molecular biologists, this approach is not practical for all researchers (Bradnam and Korf 2012). Our newly generated data is available in two ways. Raw reads and identified variants have been deposited in public databases, so that they may be accessed or incorporated into automated pipelines. NCBI has recently begun to advantageously incorporate RNA-seq data from the short-read archive into their RefSeq annotation pipeline, and the inclusion of additional unique tissue types is essential for robust annotation from this automated analysis. However, these updated annotations (especially computational predictions) are not always readily accessible in popular genome browsers. Therefore, we have also provided a downloadable BED track of our assembly. The BED format is small and much easier to use than the raw sequencing data itself. BED files also are quite easy for individual researchers to load gene model annotation into their browser of choice (Dreszer *et al*. 2011).

Our data also includes potential non-coding RNAs, which are an emerging field of research. As the RefSeq set is specifically designed for protein-coding genes,

all other transcript types are not given accession numbers. There are existing databases of non-coding RNAs available for the human and mouse genomes, however for all other species, there are only the few (less than ten) entries manually curated from the literature (Amaral *et al*. 2011). Unlike protein-coding genes, there is considerably less sequence conservation between species in non-coding RNAs, necessitating within species identification (Qu and Adelson 2012).

The function of non-coding RNAs has been the subject of recent controversy. It is debatable whether the observed RNA transcription is biologically relevant, or if transcription may simply be technical noise (van Bakel *et al*. 2010, Kapranov and Laurent 2012). Well documented functions for non-coding RNA include regulation of the genome (through chromatin modification, DNA binding, and protein binding) and of cellular differentiation during development (Rinn and Chang 2012, Hu *et al*. 2012, Morán *et al*. 2012). One of the most well-known non-coding RNAs is *XIST*, which regulates X chromosome inactivation in females. More recently, several mutations that cause overexpression of a conserved long non-coding RNA proved to be responsible for the bovine polled phenotype (Allais-Bonnet *et al*. 2013). It is thus important to consider all possible RNAs in studies of differential expression, instead of only the protein-coding transcripts.

Utilization of this data in studies of laminitis could identify new targets and pathways to help further our understanding of the etiology. Whereas current veterinary methods generally can only detect laminitis at the onset of lameness, the development of biomarkers could allow for rapid identification (and thus the most effective treatment) of cases before permanent damage occurs. Future understanding of the precise pathways underlying laminitis could lead to vital novel prevention methods and treatments.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Laura Riggs for providing the chronic samples for use in this study, as well as the owners who donated their horses to the Cornell Veterinary Hospital necropsy unit. They would also like to thank Dr. Hannah Galantino-Homer for her discussions on disease pathology and Dr. David Adelson for his computational insight as well as for providing more comprehensive equine repeat annotation.

REFERENCES

Allais-Bonnet A, Grohs C, Medugorac I, Krebs S, Djari A, Graf A, Fritz S, Seichter D, Baur A, Russ I, Bouet S, Rothammer S, Wahlberg P, Esquerré D, Hoze C, Boussaha M, Weiss B, Thépot D, Fouilloux MN, Rossignol MN, van Marle-Köster E, Hreiðarsdóttir GE, Barbey S, Dozias D, Cobo E, Reversé P, Catros O, Marchand JL, Soulas P, Roy P, Marquant-Leguienne B, Le Bourhis D, Clément L, Salas-Cortes L, Venot E, Pannetier M, Phocas F, Klopp C, Rocha D, Fouchet M, Journaux L, Bernard-Capel C, Ponsart C, Eggen A, Blum H, Gallard Y, Boichard D, Pailhoux E, Capitan A (2013) Novel insights into the bovine polled phenotype and horn ontogenesis in Bovidae. PLoS One. 8: e63512.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS (2011) lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 39: D146-151.

Bradnam K and Korf I (2012) Unix and Perl to the RESCUE! A field guide for the life sciences (and other data-rich pursuits). Cambridge University Press.

Budak MT, Orsini JA, Pollitt CC, Rubinstein NA (2009) Gene expression in the lamellar dermis-epidermis during the developmental phase of carbohydrate overload-induced laminitis in the horse. Vet Immunol Immunopathol 131: 86-96.

Capomaccio S, Vitulo N, Verini-Supplizi A, Barcaccia G, Albiero A, D'Angelo M, Campagna D, Valle G, Felicetti M, Silvestrelli M, Cappelli K (2013) RNA sequencing of the exercise transcriptome in equine athletes. PLoS One 8: e83504.

Coleman SJ, Zeng Z, Wang K, Luo S, Khrebtukova I, Mienaltowski MJ, Schroth GP, Liu J, MacLeod JN (2010) Structural annotation of equine protein-coding genes determined by mRNA sequencing. Anim Genet 41 Suppl 2: 121-130.

Coleman SJ, Zeng Z, Hestand MS, Liu J, Macleod JN (2013) Analysis of unannotated equine transcripts identified by mRNA sequencing. PLoS One 8: e70125.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W (2012) The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 40: D918-923.

Eades SC (2010) Overview of current laminitis research. Vet Clin North Am Equine Pract 26: 51-63.

Faleiros RR, Johnson PJ, Nuovo GJ, Messer NT, Black SJ, Belknap JK (2011) Laminar leukocyte accumulation in horses with carbohydrate overload-induced laminitis. J Vet Intern Med 25: 107-115.

Friedman N, Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N, Federica DP, Birren BW, Nusbaum C, Kerstin LT, Regev

A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-U130.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8: 195-202.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8: 1494-1512.

Hu W, Alvarez-Dominguez JR, Lodish HF (2012) Regulation of mammalian cell differentiation by long non-coding RNAs. EMBO Rep 13: 971-983.

Iqbal K, Chitwood JL, Meyers-Brown GA, Roser JF, Ross PJ (2014) RNA-Seq Transcriptome Profiling of Equine Inner Cell Mass and Trophectoderm. Biol Reprod 90: 61.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogentic and Genome Research 110: 462-467.

Kapranov P, St Laurent G (2012) Dark Matter RNA: Existence, Function, and Controversy. Front Genet 3: 60.

Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

Kwon S, Moore JN, Robertson TP, Hurley DJ, Wagner B, Vandenplas ML (2013) Disparate effects of LPS infusion and carbohydrate overload on inflammatory gene expression in equine laminae. Vet Immunol Immunopathol 155: 1-8.

Leise BS, Faleiros RR, Watts M, Johnson PJ, Black SJ, Belknap JK (2011) Laminar inflammatory gene expression in the carbohydrate overload model of equine laminitis. Equine Vet J 43: 54-61.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9: 34.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.

Mooney M, Bond J, Monks N, Eugster E, Cherba D, Berlinski P, Kamerling S, Marotti K, Simpson H, Rusk T, Tembe W, Legendre C, Benson H, Liang W, Webb CP (2013) Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. PLoS One 8: e61088.

Morán I, Akerman I, van de Bunt M, Xie R, Benazra M, Nammo T, Arnes L, Nakić N, García-Hurtado J, Rodríguez-Seguí S, Pasquali L, Sauty-Colace C, Beucher A, Scharfmann R, van Arensbergen J, Johnson PR, Berry A, Lee C, Harkins T, Gmyr V, Pattou F, Kerr-Conte J, Piemonti L, Berney T, Hanley N, Gloyn AL, Sussel L, Langman L, Brayman KL, Sander M, McCarthy MI, Ravassard P, Ferrer J (2012) Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metab 16: 435-448.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25: 2607-2608.

NAHMS (2001) National Economic Cost of Equine Lameness, Colic, and Equine Protozoal Myeloencephalitis in the United States. USDA:APHIS:VS, Fort Collins, CO.

The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 18, The Reference Sequence (RefSeq) Project. Available from http://www.ncbi.nlm.nih.gov/books/NBK21091

Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res 40: 10084-10097.

Noschka E, Vandenplas ML, Hurley DJ, Moore JN (2009) Temporal aspects of laminar gene expression during the developmental stages of equine laminitis. Vet Immunol Immunopathol 129: 242-253.

Park KD, Park J, Ko J, Kim BC, Kim HS, Ahn K, Do KT, Choi H, Kim HM, Song S, Lee S, Jho S, Kong HS, Yang YM, Jhun BH, Kim C, Kim TH, Hwang S, Bhak J, Lee HK, Cho BW (2012) Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. BMC Genomics 13: 473.

Pollitt CC (2010) The anatomy and physiology of the suspensory apparatus of the distal phalanx. Vet Clin North Am Equine Pract 26: 29-49.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, Dicuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42: D756-D763.

Qu Z, Adelson DL (2012) Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. PLoS One 7: e52275.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81: 145-166.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

Smit AFA, Hubley R, Green P (2010) RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.

Steelman SM, Johnson D, Wagner B, Stokes A, Chowdhary BP (2013) Cellular and humoral immunity in chronic equine laminitis. Vet Immunol Immunopathol 153: 217-226.

't Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36: e141.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol 8: e1000371.

van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43: 11.10.1-11.10.33.

van Eps AW, Pollitt CC, Underwood C, Medina-Torres CE, Goodwin WA, Belknap JK (2013) Continuous digital hypothermia initiated after the onset of lameness prevents lamellar failure in the oligofructose laminitis model. Equine Vet J. doi: 10.1111/evj.12180. [Epub ahead of print]

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody

MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867.

Wang L, Pawlak EA, Johnson PJ, Belknap JK, Eades S, Stack S, Cousin H, Black SJ (2013) Impact of laminitis on the canonical Wnt signaling pathway in basal epithelial cells of the equine digital laminae. PLoS One. 8: e56025.

Wang L, Pawlak EA, Johnson PJ, Belknap JK, Alfandari D, Black SJ (2014) Expression and activity of collagenases in the digital laminae of horses with carbohydrate overload-induced acute laminitis. J Vet Intern Med 28: 215-222.

Wilson LO, Spriggs A, Taylor JM, Fahrer AM (2014) A novel splicing outcome reveals more than 2000 new mammalian protein isoforms. Bioinformatics 30: 151-156.

Wylie CE, Collins SN, Verheyen KL, Newton JR (2012) Risk factors for equine laminitis: a systematic review with quality appraisal of published evidence. Vet J 193: 58-66.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9: e78644.

Zheng Y, Zhao LJ, Gao JP, Fei ZJ (2011) iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences. BMC Bioinformatics 12: 453.

CHAPTER 4

VARIANT IN THE *RFWD3* GENE ASSOCIATED WITH *PATN1*, A MODIFIER OF LEOPARD COMPLEX SPOTTING

H. M. Holl[1], S. A. Brooks[2], S. Archer[3], K. Brown[4], E. Bailey[5], and R. R. Bellone[4*]

[1]Department of Animal Science, Cornell University, Ithaca, NY 14853, USA

[2]Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

[3]Quill Lake, Saskatchewan S0A3E0, Canada

[4]Department of Biology, University of Tampa, Tampa, FL 33606, USA

[5]Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA

[*]Corresponding author

ABSTRACT

Leopard Complex spotting (*LP*), the result of an incompletely dominant mutation in *TRPM1*, produces a collection of unique pigmentation patterns in the domestic horse. While the *LP* mutation allows for expression of the various patterns, other loci are responsible for modification of the extent of white. Pedigree analysis of families segregating for high levels of patterning (80-100% white hair coat) indicated a single dominant gene (*PATN1*) as a major effect modifier for *LP*. Linkage analysis identified a 16 Mbp region on ECA3p associated with *PATN1*. Whole transcriptome sequencing of skin samples from horses with and without the *PATN1* allele was performed to identify genic SNPs for fine mapping. Two Sequenom assays were utilized to genotype 192 individuals from five *LP*-containing breeds. The initial panel highlighted a 1.6 Mbp region without a clear candidate gene. In the second round of fine-mapping, a SNP in the 3' UTR of *RING finger and WD repeat domain 3* (*RFWD3*) reached a significance level of $p=1.063x10^{-39}$. Sequencing of *RFWD3* did not identify any coding polymorphisms specific to the *PATN1* horses. Genotyping of an additional 52 *LP* animals increased this association to a p-value of $4.68x10^{-56}$. An additional 166 horses of breeds not segregating for *LP* did not contain the *PATN1*-associated allele. This variant is a strong candidate for *PATN1* and may be used for genotyping *lp*/*lp* animals.

INTRODUCTION

Coat color is a highly valued trait in domestic species. Studies of ancient DNA have suggested that the wide variation of pigmentation seen in the horse is partly due to human artificial selection throughout the process of domestication (Ludwig *et al*. 2009). The relative ease of phenotyping makes coat color an ideal model system. Genetic mapping studies have been quite successful in the horse, leading to the availability of a variety of molecular tests that are routinely used by breeders (Rider 2009).

Most coat colors are simple in inheritance, with a single autosomal allele responsible for a distinct alteration in phenotype. Leopard complex spotting, however, is a collection of related patterns found in several breeds with a more complex inheritance (Sponenberg *et al*. 1990). A single incompletely dominant allele (*LP*) allows for the expression of a range of pigmentation phenotypes, including variable symmetrical white patterning centered over the hips, striped hooves, white sclera, mottled skin, and progressive depigmentation of the hair known as “roaning” (Bellone *et al*. 2013). Heterozygotes generally have many oval pigmentation spots (1-5 cm) within regions of white coat, whereas homozygotes have few to no such spots. The wide range of variation seen in the extent of white patterning is believed to be due to other modifying loci.

Previous research has linked *LP* to a retroviral insertion into an intron of *transient receptor potential cation channel subfamily M member 1* (*TRPM1*), resulting in premature polyadenylation (Bellone *et al*. 2013). TRPM1 is a calcium channel present in the brain, heart, melanocytes, and retina. The function of *TRPM1* in most tissues is not well understood, though studies in the retina have thoroughly investigated its involvement in low-light vision, as reviewed in Oancea and Wicks 2011. Horses homozygous for *LP* are affected by congenital stationary night blindness

(CSNB), consistent with a causal role of the mutation in *TRPM1*, though the molecular mechanism responsible for the pigmentation phenotype has yet to be elucidated (Sandmeyer *et al*. 2007, Sandmeyer *et al*. 2012).

Variability in white pattern expression is not a unique observation to leopard complex spotting. Studies of white markings in the Arabian horse revealed that there are both genetic and stoichastic factors that influence pattern development, and respond readily to selection (Rieder *et al*. 2008). Notably, the *e* allele of the *MC1R* gene (which leads inability to produce eumelanin) and the *A* allele of the *ASIP* gene (eumelanin production restricted to the extremities) are associated with increased size of white markings. Male horses were also shown to have larger markings than female horses. This observation has been made in other breeds and patterns, including in leopard complex spotting (Sponenburg 2003). Recent work in the Franches-Montagnes horse have shown an accumulation of mutations in *KIT* and *MITF* are responsible for an increase in white markings (Haase *et al*. 2013).

However, there are some modifying loci that appear to be unique to *LP*. One early observation by The Appaloosa Project (http://www.appaloosaproject.info/) researchers was that horses with the more extreme white patterns (termed leopard and few-spot) always had a parent with these same patterns (Sponenburg 2009). Breeding records suggested this was a single autosomal dominant allele, possibly displaying incomplete dominance. The effects of this allele, named *PATN1* for *pattern-1*, are shown in Figure 4.1. Crosses to breeds without leopard complex spotting indicated that this phenotype was only found in breeds segregating for *LP*, and that it had no discernible effect on other white patterning.

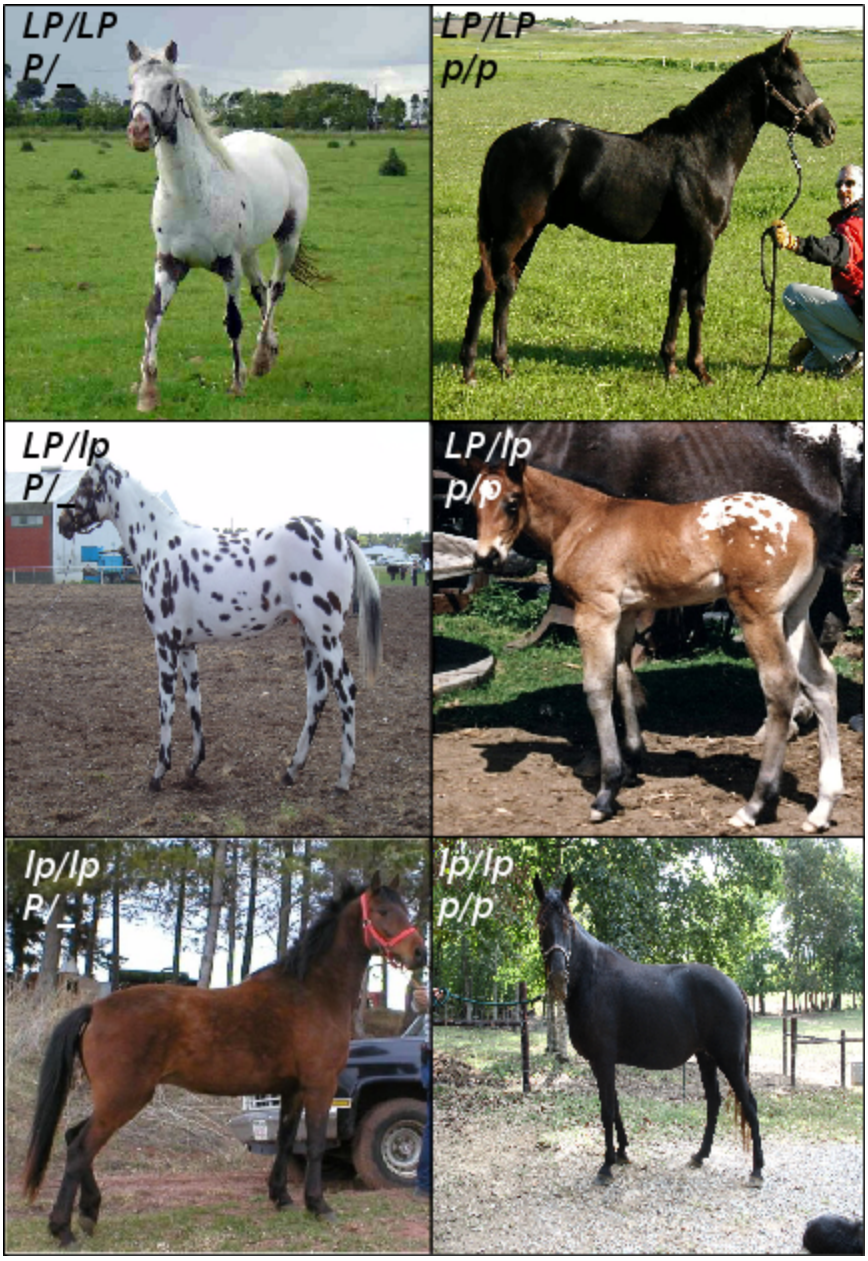

Figure 4.1. Phenotypes for zygosity at *LP* and *PATN1*. *LP*, which acts in an incompletely dominant fashion, enables the expression of *PATN1*.

Additional pedigree studies identified an association between *PATN1* and *MC1R* alleles, suggesting linkage and thus localization to ECA3. In this study, we first used linkage analysis in two half sibling families to map *PATN1* onto ECA3, then used whole transcriptome sequencing to generate variants for fine-mapping. Although the implicated region had no clear candidate genes or mutations, further genotyping suggested a 3'UTR variant of *RFWD3* as closely linked to the *PATN1* allele. Here we present the results of mapping *PATN1* and the analysis of *RFWD3*.

MATERIALS AND METHODS

Sample collection and phenotyping

Two half-sibling families were available for linkage mapping, consisting of few-spot Appaloosa stallions (*LP*/*LP PATN1*/*patn1*) mated to Thoroughbred and American Quarter Horse mares (breeds not known to have *PATN1*). Family A consisted of 17 offspring (9 *PATN1* and 9 *patn1*) and family B consisted of 30 offspring (13 *PATN1* and 17 *patn1*). Only offspring with clear phenotypes at birth were included (>70% white for *PATN1*, <30% white for *patn1*, as described below).

Blood or hair samples were collected for genotyping from 245 horses with *LP* (117 Appaloosas [ApHC], 13 British Spotted Ponies [BSP], 87 Knabstruppers [KB], 25 Miniature Horses [MINI], 2 Pony of Americas [PoA]). Of these, 192 were used for fine-mapping (92 ApHC, 10 BSP, 69 KB, 19 MINI, 2 PoA). An additional 166 horses were available from banked DNA of breeds that are not believed to carry *PATN1* (51 Arabians, 30 Standardbreds, 32 Thoroughbreds, 51 Quarter Horses, and 2 Arabian/Quarter Horse crosses). DNA from blood samples was extracted either using the Puregene whole-blood extraction kit (Qiagen Inc., Valencia, CA, USA) or Nucleon Bacc2 kit (GE Healthcare Bio-Sciences Corp., Piscataway, NJ, USA). High-quality DNA from hair was prepared for Sequenom genotyping by a modified Puregene

protocol (Cook *et al*. 2010). Hair lysates were prepared for restriction digest genotyping (Locke *et al*. 2003).

A single experienced phenotyper determined the presence of the *PATN1* allele using photographic records, parental phenotypes, and progeny records. The *PATN1* phenotype was defined by at least 60% white coat at birth and at least one parent with production records consistent with the allele. In order to control for small effect modifiers, the *patn1* phenotype was assigned to horses with less than 40% white. Horses with large white face and leg markings, consistent with the presence of another white patterning allele, were excluded, as these are known to affect leopard complex patterning (Hauswirth *et al*. 2013).

Linkage mapping of ECA3

Nine microsatellites from ECA3 were chosen for analysis. An ABI 310 genetic analyzer was used to detect fluorescently labeled primers (Applied Biosystems Inc., Foster City, CA, USA). The STRand microsatellite analysis software (available at http://www.vgl.ucdavis.edu/informatics/STRand/) was used to analyze the data (Toonen and Hughes 2001). Fisher's exact tests were performed using 2x2 contingency tables.

RNA extraction and sequencing

Full thickness skin biopsies were obtained from one Appaloosa (*PATN1*/*PATN1*) and one Thoroughbred (*patn1*/*patn1*). Punch biopsies were performed as previously described (Bellone *et al*. 2008). Total RNA was isolated either using the RNeasy Lipid Tissue mini kit (Qiagen Inc., Valencia, CA, USA) according to the manufacturer's protocol or by acid guanidinium thiocyanate-phenol-chloroform extraction (Chomczynski and Sacchi 1987, MacLeod *et al*. 1996). Quantification was

performed using a NanoDrop spectrophotometer (NanoDrop Technologies LLC., Wilmington, DE, USA) and quality was assessed using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Library preparation and sequencing was performed by Cornell University's Life Sciences Core Laboratory Center. Single-end libraries were constructed using manufacturer's protocols for poly-T selection and sequenced on an Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA). The resulting reads were aligned to the EquCab2 reference genome using the BWA software package under default parameters (Wade *et al*. 2009, Li and Durbin 2009). SAMtools was used to convert alignments to BAM format for visualization and to call variants (Li *et al*. 2009). Visualization was performed using the UCSC Genome Browser and IGV (Kent *et al*. 2002, Robinson *et al*. 2011). Raw reads were submitted to the European Nucleotide Archive (accession number PRJEB6101).

Sequenom SNP genotyping

Variants were called using SAMtools “mpileup” and were filtered to retain those with a phred-scaled quality over 30 and at least four observations. RNA-seq data from four additional horses (two *patn1/patn1* two unknown) were available from a previous study (Bellone *et al*. 2013). Identified variants were compared to create two pools of SNPs for fine-mapping. The first were homozygous SNPs that were observed only in the single *PATN1*/*PATN1* sample. The second set was comprised of SNPs that were observed in at least two, but not all sequenced horses, corresponding to polymorphic loci in the Appaloosa breed. Variants from areas without RNA-seq coverage were obtained from the published SNP database (Wade *et al*. 2009).

Two fine-mapping sets of 80 SNPs were generated for genotyping with the Sequenom Mass Spectrophotometry platform using the iPlex system (Gabriel *et al*.

2009). Multiplex design and assays were carried out by Geneseek (Geneseek Inc. Lincoln, NE, USA). An initial panel spanned the 16 Mbp region with the strongest association in the linkage data, with preference given to include the *PATN1*-specific variants in the assay. A second panel spanning the refined 1.6 Mbp region contained the top 5 markers from the first panel and additional SNPs representing almost all 34 genes in the region (Table 4.1).

Genotyping files were analyzed using PLINK 1.07 (Purcell *et al*. 2007). Filtering parameters were 85% individual call rate, 90% genotyping rate, and 5% minor allele frequency. After filtering, there were 186 individuals and 74 SNPs in the first panel and 192 individuals and 68 SNPs in the second panel. Association was performed using the genotypic model with the Fisher's exact test. Analyses were performed on all breeds, only Appaloosas, and only Knabstruppers. Haploview was used to assess haplotypes and linkage disequilibrium (Barrett *et al*. 2005).

Table 4.1. Summary of SNPs used in Sequenom assay design. *PATN1*-specific SNPs were only detected in *PATN1* sequences (derived from a single individual) whereas *LP*-polymorphic SNPs were found in multiple *LP* sample sequences. Available samples were of genotypes *LP*/*LP PATN1*/_, *LP*/*LP patn1*/*patn1*, and *LP*/*lp patn1*/*patn1*.

Set	Total	Panel 1	Panel 2
PATN1-specific	100	25	11
LP-polymorphic	288	19	34
EquCab2	330	36	35
Total	718	80	80

RT-PCR and sequencing of candidate gene RFWD3

Transcripts assembled from hoof RNA-seq in an unpublished study were aligned to equCab2 using BLAT and visualized with the UCSC Genome Browser (Kent 2002). The most associated fine-mapping SNP fell within a transcript corresponding to the 3'UTR of *RING finger and WD repeat domain 3* (*RFWD3*). Coverage of mapped skin RNA-seq reads was used to incorporate the reference DNA sequence into the transcript, expanding on a missing section of the 5'UTR. All samples with defined *PATN1* genotypes were then aligned to the *RFWD3* transcript and loaded into IGV. Alignments were visually inspected to confirm full coverage of all exons in the model, as well as to identify any additional polymorphisms.

Due to inconsistent coverage of RNA-seq reads, Sanger sequencing was selected for closer examination of *RFWD*3. Primers spanning exon 10, both UTRs, and a segment from exon 9 to the 3'UTR were designed based on the transcript alignment to EquCab2 using the Primer3 software (Table 4.3, Rozen and Skaletsky 1998). Two-step RT-PCR was performed using the SuperScript VILO MasterMix kit (LifeTechnologies, Carlsbad, CA, USA) followed by standard PCR. All DNA was amplified using 20 μL volume PCR with FastStart Taq DNA polymerase (Roche Applied Science, Branford, CT, USA) and included all reagents per the manufacturers recommended conditions.

PCR products were submitted to the Cornell Core Life Sciences Laboratories Center for sequencing using standard ABI chemistry on a 3730 DNA Analyzer (Applied Biosystems Inc., Foster City, CA, USA). Amplicons were aligned and screened for mutations using Consed (Gordon *et al*. 1998).

Table 4.2. Primers used to verify RNA-seq observations. All PCRs were performed with an elongation time of 30s.

Name	Forward Seq	Reverse Seq	Size	T
RFWD3-3'UTR-1	AAGAGGCAGACCTGTTCCTG	TAGGAACCGAAGGAAGCTGA	455 bp	62°C
RFWD3-3'UTR-2	CAGGTCATAACTGTTGTAAAATAAAAATAT	TATTTATTCTCAGGAAGATAATAGTGGTTC	311 bp	55°C
RFWD3-5'UTR	TGTCCTGTGTGAAGCAGACC	GCTTTCCCAGTAATGGCTCA	801 bp	62°C
RFWD3-exon10	CTATGTCCCCGTTCACACAA	CTTGAGGGCTCCTGGTAGTG	492 bp	62°C
RFWD3-ex9-3'U	CCATGGTCACTCCCACCTAT	CTGCTGCTGGTGTCTTGATG	700 bp	62°C

PCR-RFLP SNP Genotyping

Two SNPs unique to the single *PATN1* horse were detected in the 3'UTR of *RFWD3*. In both cases, genotyping was performed by 10 μL PCR followed by 20 μL restriction digestion and visualization on 4% agarose gels. SNP ECA3:23,658,447T>G (RFWD3-3U1) was digested for 20 minutes at 37°C using EcoRV (New England BioLabs Inc., Ipswitch, MA, USA), resulting in either a single 455bp band representing the G allele, or 276bp+170bp bands for the T allele. As SNP ECA3:23,659,162T>C (RFWD3-3U2) does not normally alter a restriction site, dCAPS Finder 2.0 was used to design PCR primers to generate a TaqI site specific to the C allele, resulting in 281bp+30bp bands, versus a single 311 bp band for the T allele (Neff *et al*. 2002). SNP genotypes were tested for association to *PATN1* by a Fisher's exact test using R 3.0.1 (R Core Team 2013).

RESULTS

Illumina RNA-seq analysis

Whole transcriptome sequencing of the three skin samples produced a total of 105,509,542 reads (Table 4.4). Visualization of the alignments in UCSC did not reveal any extreme differences in gene expression within the *PATN1* associated region. As no replicates were available, formal differential expression analysis with software packages was not attempted.

Table 4.3. Sample information and statistics from whole transcriptome sequencing of full-thickness skin.

Sample	Phenotype	Total Reads	Mapped Reads	% Mapped	SNPs
06-92 pigmented	*PATN1*	33,133,824	29,094,811	88%	85,027
06-92 unpigmented	*PATN1*	30,292,109	26,369,281	87%	77,711
D-052 pigmented	*patn1*	41,083,268	34,312,745	84%	45,178

Sequenom fine-mapping

Linkage mapping using 9 microsatellite markers on ECA3 was undertaken in two half sibling families segregating for *PATN1*. The strongest association signal ranged from 11 Mbp to 18.2 Mbp (Table 4.4). Due to the observed gene density, SNP selection for fine-mapping was expanded out to 26 Mbp.

The initial fine-mapping panel showed a strong association ($p<2.4x10^{-18}$) from ECA3:22,043,005-23,677,469 (Figure 4.2A, best hit SNP ECA3:23,677,469G>A, $p=3.539x10^{-23}$). None of the 34 genes in this region possess a previously reported function in pigmentation or a coat color phenotype. Association signal strengthened in the second set of markers (Figure 4.2B, top hit ECA3:23,658,447T>G, $p=1.063x10^{-39}$). This SNP lies in the 3' UTR of *RFWD3*. There were no differences in the position of the primary association signal when analyzing the dataset by breed (Appaloosa $p=2.288x10^{-25}$, Knabstrupper $p= 4.359x10^{-16}$).

Although there was strong LD present in this region, haplotype analysis suggested that in this population there is little additional recombination left for further refining the association signal. Significantly associated haplotypes were defined by the presence of single SNPs with strong associations from the fine-mapping and did not provide any additional information. Recombination between the top SNP and its nearest neighbors prevented additional analysis.

Table 4.4. Statistics from microsatellite linkage mapping. Marker order and linkage position are derived from Penedo *et al*. 2005. P-values were generated by Fisher's Exact Test on 2x2 contingency tables. ND = not determined

Marker	Linkage Position	EquCab2 Position	Family A	Family B
AHT036	0	29.4 Mb	0.30419	0.25975
COR028	12.5 cM	11.1 Mb	0.00029	0.00020
AHT022	20.4 cM	21.1 Mb	0.00108	ND
TKY215	28.3 cM	18.2 Mb	ND	0.00001
UCDEQ437	30.7 cM	31.3 Mb	ND	0.00184
TKY651	39.9 cM	37.9 Mb	ND	0.00278
TKY528	39.9 cM	39.8 Mb	0.04004	0.00031
TKY447	55.0 cM	70.0 Mb	ND	0.03097
ASB023	59.1 cM	79.2 Mb	0.04446	0.16285

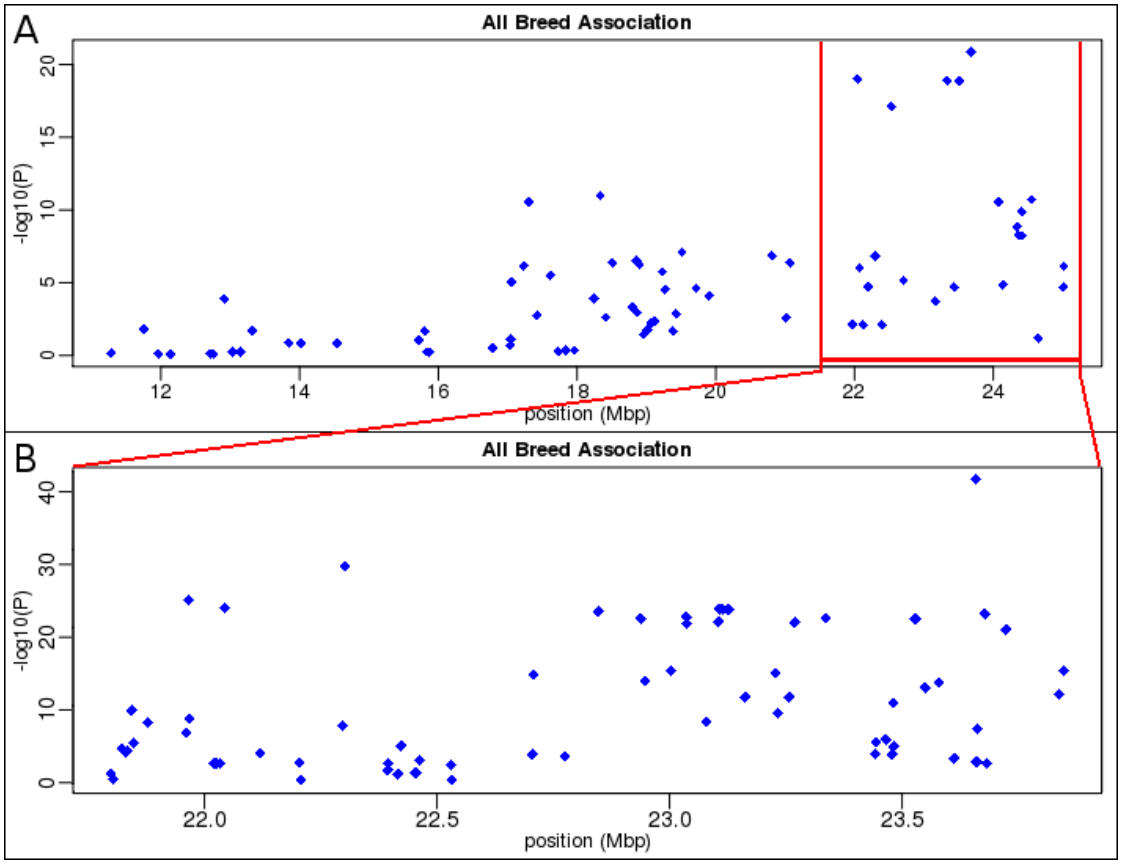

Figure 4.2. Manhattan plots for fine-mapping of *PATN1*. **A** The initial 16 Mbp panel showed a moderate association from ECA3:22M-24M. The red line indicates the region selected for the second round of fine-mapping **B** The 1.6 Mbp panel displayed a similar association signal to the first, with a new SNP showing a strong association.

Analysis of RFWD3

The gene *RFWD3* was selected for further analysis on the basis of the strong association signal and recombination seen between all other genotyped SNPs. The next closest gene, *GLG1*, was not a likely candidate because it did not possess variants specific to the *PATN1* horse in the RNA-seq data. Alignments to the hoof *RFWD3* transcript showed full coverage with no reads indicative of missing exons or alternative splicing (Figure 4.3). There was a dip of coverage over exon 10 apparent in the *PATN1* horse (from 70 read depth to 15 read depth) that was not observed in the other two samples. Coverage over the 5'UTR was significantly lower in all animals, and closer examination of the reference sequence showed highly similar 51 bp repeats present in both the reference and assembled hoof data. One additional variant in the 3'UTR and two in the 5'UTR were the only markers uniquely observed in the *PATN1* horse. However, Sanger sequencing did not reveal any additional polymorphisms, with all amplicons producing high-quality sequences of the expected length. The abnormal coverage from RNA-seq was likely due to computational artifacts.

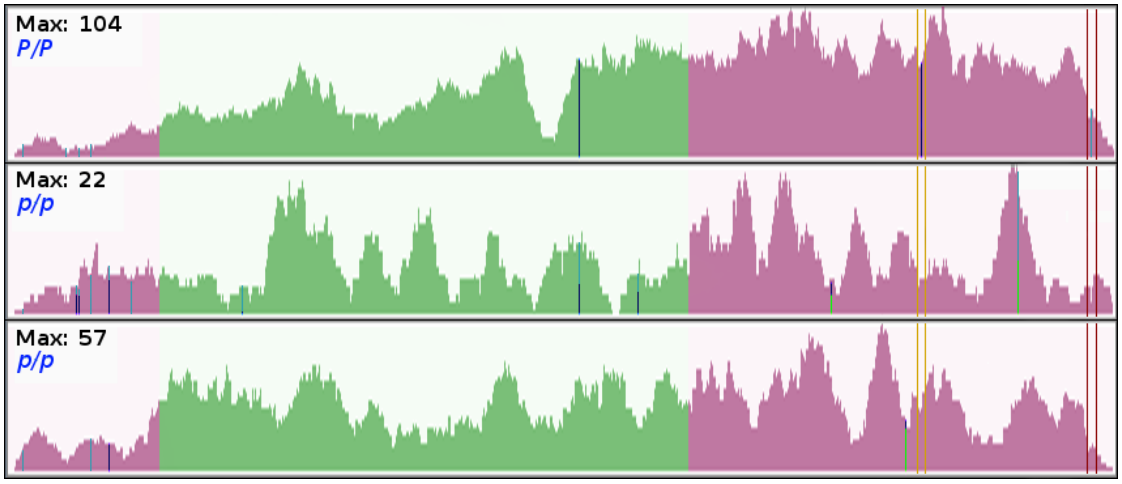

Figure 4.3. Coverage of RNA-seq reads to hoof-derived *RFWD3* transcript assembly. The purple sections are the 5' and 3' untranslated regions, whereas the green section is the open reading frame. Vertical lines indicate a variant from the reference. The max value is the maximum read depth across the alignment. SNP RFWD3-3U1 is outlined in red and SNP RFWD3-3U2 is in orange.

Genotyping of additional animals

An additional 218 horses were genotyped for SNP RFWD3-3U1 using PCR-RFLP. 52 of these horses were from *LP*-containing breeds and thus had been phenotyped for *PATN1*. 166 horses were from non-*LP* breeds and thus were not expected to carry *PATN1*. Data from the Sequenom panel and PCR-RFLP genotyping were pooled for statistical analysis (Table 4.5). SNP RFWD3-3U1 was strongly associated with *PATN1* (p=4.68e-56) and was not found in any of the non-*LP* breed samples.

Genotypes for SNP RFWD3-3U2 were determined in 39 horses with definite *PATN1* phenotypes. Although 35 horses showed T-T or C-G linkage with the RFWD3-3U1 SNP, there were four horses with a recombinant C-T haplotype (Table 4.6). As all of these horses exhibited the PATN1 phenotype perfectly linked to SNP RFWD3-3U1, additional horses were not genotyped for RFWD3-3U2.

Table 4.5. Fisher's exact tests for RFWD3-3U1. *LP*-containing breeds are labeled ApHC (Appaloosas), BSP (British Spotted Ponies), KB (Knabstruppers), MINI (Miniature Horses), and PoA (Pony of Americas). The Non-*LP* set was comprised of 51 Arabians, 30 Standardbreds, 32 Thoroughbreds, 51 Quarter Horses, and 2 Arabian/Quarter Horse crosses.

Breed	**Pheno**	**G/_**	**T/T**	**Total**	**P-value**
ApHC	*PATN1*	55	2	57	
	patn1	0	60	60	
	Total	55	62	117	1.61e-31
KB	*PATN1*	58	1	59	
	patn1	1	27	28	
	Total	59	28	87	3.32e-20
MINI	*PATN1*	6	0	6	
	patn1	5	14	19	
	Total	11	14	25	n/a
BSP	*PATN1*	9	0	9	
	patn1	1	3	4	
	Total	10	3	13	n/a
PoA	*PATN1*	1	0	1	
	patn1	0	1	1	
	Total	1	1	2	n/a
Total *LP*	*PATN1*	129	3	132	
	patn1	7	105	112	
	Total	136	108	244	4.68e-56
Non-*LP*	*PATN1*	0	0	0	
	patn1	0	166	166	
All breed	p-value				1.55e-93

Table 4.6. Linkage between SNPs RFWD3-3U1 and RFWD3-3U2. Genotypes are listed as [RFWD3-3U1]-[RFWD3-3U2], with the top section demonstrating the more common linkage (G-C and T-T) and the bottom section demonstrating recombination (T-C). The two SNPs are located 715 bp apart. Each column lists the number of observations of each linkage type by breed (ApHC [Appaloosa], BSP [British Spotted Pony], KB [Knabstrupper], MINI [Miniature Horse], PoA [Pony of Americas]).

3' UTR SNPs (U1-U2)	ApHC	KB	MINI	BSP	POA	Total
GG-CC	3	5	0	1	1	10
GT-CT	8	2	2	3	0	15
TT-TT	7	1	2	0	0	10
GT-CC	0	0	1	1	0	2
TT-CT	1	0	0	0	1	2

DISCUSSION

Leopard complex spotting is a collection of coat color patterns resulting from the incompletely dominant *LP* allele of *TRPM1* and its interactions with other modifying loci. One such genetic factor, *PATN1*, is responsible for large differences in pattern levels segregating in observed families. Linkage between *PATN1* and *MC1R* alleles as well as with microsatellite markers on ECA3 support localization to that chromosome. Whole transcriptome sequencing and two rounds of fine mapping by targeted association identified a polymorphism in the 3' UTR of *RFWD3* with strong association with *PATN1*. This variant was not detected in any of the four non-*LP* breeds tested (three of which were used in the formation the Appaloosa breed). Although the variant was not in complete concordance with the phenotype, there was no other detected variant with a stronger association.

Although clear differences in siblings segregating for the *PATN1* allele can be easily observed, the complex nature of white patterning can make phenotyping of unrelated individuals difficult. Beyond known factors such as *MC1R* allele and sex, there may exist many unknown genes involved in the enhancement or repression of white patterning. Variation in the degree of depigmentation comprising white patterning was previously documented in both the overo and tobiano coat colors (Lightbody 2002, Stamatelakys 2011). Despite strong functional evidence and similarity to homologous phenotypes tying these patterns to their respective mutations, individuals with extremely low levels of white patterning were described. The most extreme examples have been reported in Miniature Horses, where individuals genotyped positive for a dominant white spotting allele, but lacked even a single discernible white hair. For example, Santschi *et al*. 2001 observed two such Miniature Horses that had genotyped heterozygous for the frame overo pattern. Five of the Miniature Horses in this study typed heterozygous for RFWD3-3U1 despite having

below 40% white patterning and thus may be further examples of this extreme white pattern suppression.

RING finger and WD repeat domain 3 (RFWD3) is an E3 ubiquitin ligase involved in DNA damage response. Initial studies indicated that in response to DNA damage, RFWD3 forms a complex with Mdm2, another E3 ubiquitin ligase that normally degrades p53 (Fu *et al*. 2010). However, the RFWD3-Mdm2 complex instead stabilizes p53, which activates the G1 checkpoint. Once this checkpoint has passed, RFWD3 becomes inactive and Mdm2 is able to return p53 to its usual low-level of expression. RFWD3 also associates with replication protein A and is recruited to sites of DNA damage (Gong and Chen 2011, Liu *et al*. 2011).

There are no *in vivo RFWD3* mutants present in the literature. However, phenotypic similarities exist between *PATN1* and harlequin, a pigmentation phenotype in the domestic dog. Harlequin Great Danes are the result of an interaction of the *M* and *H* alleles. *M* has been identified as a SINE insertion in the *SILV* gene, resulting in a coat dilution phenotype with incomplete dominant inheritance (Clark *et al*. 2006). Spontaneous mutations within the insertion result in variable patches of fully pigmented coat. The lethal-dominant *H* allele selectively removes the dilute color, leading to no visible phenotype in *m/m* individuals and fully pigmented patches on a completely depigmented background on dogs possessing the *M* allele. *H* is the result of a missense mutation in *PSMB7*, which codes for the β2 catalytic subunit of the proteasome (Clark *et al*. 2011). The authors speculate that mutant transcripts of *M* may generate abnormal proteins targeted for degredation by the ubiquitin proteasome pathway. However, as the proteasomes in *H* animals are unable to effectively clear these proteins, there is either an accumulation of cytotoxic components leading to melanocyte death or abnormal $SILV^{M}$-$SILV^{m}$ protein complexes that alter cell morphology. The resulting patches containing the abnormal protein thus lack normal

melanocytes and are unable to produce pigment, leading to the shift from merle coloration to white in the harlequin dogs.

The depigmentation phenotype seen in the *LP+PATN1* horse could be due to a similar interaction. One theory is that abnormal melanosome morphology, observed in *LP* fibroblast cultures, results in a death of melanocytes, leading to regions lacking pigmentation at birth, as well as a progressive roaning through life (Bellone *et al*. 2013). One depigmentation phenotype in zebrafish, caused by a mutation in *trpm7*, was shown to be the result of abnormal melanosome morphology leading ruptured cellular membranes and melanophore death (McNeill *et al*. 2007). Inhibition of melanin synthesis was shown to rescue the phenotype, implicating a mechanism of an accumulation of cytotoxic intermediates. If *RFWD3* is involved in the removal of deformed melanosomes seen in *LP* horses, either directly or through downstream targets, alterations of its expression could lead to a much earlier melanocyte death. The spots of pigmentation seen in *LP*/*lp* horses could be the result of populations of melanocytes with sufficient functional *TRPM1* to avoid cytotoxicity throughout life.

The observed 3'UTR polymorphism was not perfectly associated to the *PATN1* phenotype. It is possible that this variant is merely in close linkage to the casual mutation, and further study is needed to screen for additional mutations. Another possibility is that the difficulty in phenotyping (due to other modifying loci) may have led to misclassification of some horses. In this case, the polymorphism could be responsible for misexpression of *RFWD3* during development, or for disruption of normal translational control through an alteration of RNA secondary structures (Jia *et al*. 2013). Structures in the 3'UTR are often targeted by proteins and miRNAs in order to moderate translation into proteins . As transcriptome sequencing did not generate markers with fine enough resolution to define haplotypes, targeted genomic re-sequencing could be used to more thoroughly examine this region. However, given the

strong association signal, genotyping of the RFWD3-3U1 (ECA3:23,658,447T>G) SNP could be used by breeders to identify *lp*/*lp* individuals that are likely to possess the *PATN1* allele, as this color is a highly desirable trait in the breed.

ACKNOWLEDGMENTS

The authors would like to thank all of the breeders and owners who collected samples from their horses for this study. We thank Taryn Cranford for her technical assistance as part of her Independent Study Courses at the University of Tampa as well as Dr. Meco Bernoco for his technical assistance. The study was supported in part by grants from the Appaloosa Horse Club of Canada and by generous donations by Appaloosa breeders who belong to the Appaloosa Project electronic classroom. Karla Brown was supported by a University of Tampa Department of Biology Summer research fellowship.

REFERENCES

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265.

Bellone RR, Brooks SA, Sandmeyer L, Murphy BA, Forsyth G, Archer S, Bailey E, Grahn B (2008) Differential gene expression of TRPM1, the potential cause of congenital stationary night blindness and coat spotting patterns (LP) in the Appaloosa horse (Equus caballus). Genetics 179: 1861-1870.

Bellone RR, Holl H, Setaluri V, Devi S, Maddodi N, Archer S, Sandmeyer L, Ludwig A, Foerster D, Pruvost M, Reissmann M, Bortfeldt R, Adelson DL, Lim SL, Nelson J, Haase B, Engensteiner M, Leeb T, Forsyth G, Mienaltowski MJ, Mahadevan P, Hofreiter M, Paijmans JL, Gonzalez-Fortes G, Grahn B, Brooks SA (2013) Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. PLoS One 8: e78280.

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162: 156-159.

Clark LA, Tsai KL, Starr AN, Nowend KL, Murphy KE (2011) A missense mutation in the 20S proteasome β2 subunit of Great Danes having harlequin coat patterning. Genomics 97: 244-248.

Clark LA, Wahl JM, Rees CA, Murphy KE (2006) Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc Natl Acad Sci U S A 103: 1376-1381.

Cook D, Gallagher PC, Bailey E (2010) Genetics of swayback in American Saddlebred horses. Anim Genet 41 Suppl 2: 64-71.

Fu X, Yucer N, Liu S, Li M, Yi P, Mu JJ, Yang T, Chu J, Jung SY, O'Malley BW, Gu W, Qin J, Wang Y (2010) RFWD3-Mdm2 ubiquitin ligase complex positively regulates p53 stability in response to DNA damage. Proc Natl Acad Sci U S A 107: 4579-4584.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP Genotyping using the sequenom MassARRAy iPLEX platform. Curr Protoc Hum Genet Chapter 2: Unit 2.12.

Gong Z, Chen J (2011) E3 ligase RFWD3 participates in replication checkpoint control. J Biol Chem 286: 22308-22313.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8: 195-202.

Haase B, Signer-Hasler H, Binns MM, Obexer-Ruff G, Hauswirth R, Bellone RR, Burger D, Rieder S, Wade CM, Leeb T (2013) Accumulating mutations in series of haplotypes at the KIT and MITF loci are major determinants of white markings in Franches-Montagnes horses. PLoS One 8: e75071.

Hauswirth R, Jude R, Haase B, Bellone RR, Archer S, Holl H, Brooks SA, Tozaki T, Penedo MC, Rieder S, Leeb T (2013) Novel variants in the KIT and PAX3 genes in horses with white-spotted coat colour phenotypes. Anim Genet 4: 763-765.

Jia J, Yao P, Arif A, Fox PL (2013) Regulation and dysregulation of 3'UTR-mediated translational control. Curr Opin Genet Dev 23: 29-34.

Kent WJ (2002) BLAT - the BLAST-like alignment tool. Genome Res 12: 656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Lightbody T (2002) Foal with Overo lethal white syndrome born to a registered quarter horse mare. Can Vet J 43: 715-717.

Liu S, Chu J, Yucer N, Leng M, Wang SY, Chen BP, Hittelman WN, Wang Y (2011) RING finger and WD repeat domain 3 (RFWD3) associates with replication protein A (RPA) and facilitates RPA-mediated DNA damage response. J Biol Chem 286: 22314-22322.

Locke MM, Penedo MC, Bricker SJ, Millon LV, Murray JD (2002) Linkage of the grey coat colour locus to microsatellites on horse chromosome 25. Anim Genet 33: 329–337.

Ludwig A, Pruvost M, Reissmann M, Benecke N, Brockmann GA, Castaños P, Cieslak M, Lippold S, Llorente L, Malaspinas AS, Slatkin M, Hofreiter M (2009) Coat color variation at the beginning of horse domestication. Science 324: 485.

MacLeod JN, Burton-Wurster N, Gu DN, Lust G (1996) Fibronectin mRNA splice variant in articular cartilage lacks bases encoding the V, III-15, and I-10 protein segments. J Biol Chem 271: 18954-18960.

McNeill MS, Paulsen J, Bonde G, Burnight E, Hsu MY, Cornell RA (2007) Cell death of melanophores in zebrafish trpm7 mutant embryos depends on melanin synthesis. J Invest Dermatol 127: 2020-2030.

Penedo MC, Millon LV, Bernoco D, Bailey E, Binns M, Cholewinski G, Ellis N, Flynn J, Gralak B, Guthrie A, Hasegawa T, Lindgren G, Lyons LA, Røed KH, Swinburne JE, Tozaki T (2005) International Equine Gene Mapping Workshop

Report: a comprehensive linkage map constructed with data from new markers and by merging four mapping resources. Cytogenet Genome Res 111: 5-15.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rieder S (2009) Molecular tests for coat colours in horses. J Anim Breed Genet 126: 415-424.

Rieder S, Hagger C, Obexer-Ruff G, Leeb T, Poncet PA (2008) Genetic analysis of white facial and leg markings in the Swiss Franches-Montagnes Horse Breed. J Hered 99: 130-136.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative Genomics Viewer. Nat Biotechnol 29: 24–26.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

Sandmeyer LS, Bellone RR, Archer S, Bauer BS, Nelson J, Forsyth G, Grahn BH (2012) Congenital stationary night blindness is associated with the leopard complex in the miniature horse. Vet Ophthalmol 15: 18-22.

Sandmeyer LS, Breaux CB, Archer S, Grahn BH (2007) Clinical and electroretinographic characteristics of congenital stationary night blindness in the Appaloosa and the association with the leopard complex. Vet Ophthalmol 10: 368-375.

Santschi EM, Vrotsos PD, Purdy AK, Mickelson JR (2001) Incidence of the endothelin receptor B mutation that causes lethal white foal syndrome in white-patterned horses. Am J Vet Res 62: 97-103.

Sponenberg DP, Carr G, Simak E, Schwink K (1990) The inheritance of the leopard complex of spotting patterns in horses. J Hered 81: 323-331.

Sponenberg DP (2003) Equine Color Genetics, 2nd edition. Wiley-Blackwell, Hoboken, NJ, USA. pp. 89.

Sponenberg DP (2009) Equine Color Genetics, 3rd edition. Wiley-Blackwell, Hoboken, NJ, USA. pp. 118.

Stamatelakys I (2011) Slipped Away. Paint Horse Journal 45: 68-75.

Toonen RJ, Hughes S (2001) Increased Throughput for Fragment Analysis on ABI Prism 377 Automated Sequencer Using a Membrane Comb and STRand Software. Biotechniques 31:1320-1324.

Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole

Genome Assembly Team, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867.

CHAPTER 5

SUMMARY

The field of equine genetics rapidly changed over the past decade. The first studies relied on information from other species, investigating candidate genes based on shared phenotypes (Rudolph *et al*. 1992, Shin *et al*. 1997, Metallinos *et al*. 1998, Santschi *et al*. 1998, Yang *et al*. 1998). The collaborative international horse genome project greatly accelerated genetics research, leading to the maps and tools that enabled the discovery of many more causal variants. Following completion of the reference genome sequence, researchers could immediately access horse-specific sequence information. However, gene annotation was still largely based on comparative genetics and computational predictions. The rise of the field of genomics resulted in a similar enhancement of research in many species, allowing for the efficient sequencing of the entire genetic material of an individual, with or without a reference sequence. In human clinical genetics, whole exome sequencing (the coding portions of the genome) is becoming increasingly common as a diagnostic tool, producing putative causal variants without the need for genetic mapping (Rabbani *et al*. 2014). A recent study applied a similar analysis in the horse, employing whole-genome sequencing and screening mutations for coding variants (Towers *et al*. 2013). The incorporation of these newer technologies is essential for the future of equine genetics.

The work presented in this dissertation is an extension of this goal, applying novel methodologies to the genetics of the horse. In Chapter Two, the Illumina SNP array was used to diagnose several chromosomal abnormalities. Microarray analysis is becoming increasingly common in humans, providing an important tool for the diagnosis of congenital anomalies and developmental delays (Henderson *et al*. 2014).

Traditional methods such as FISH work well for identifying issues with the larger chromosomes and translocations, but rely on cell culture and thus high quality tissue samples. One advantage of the SNP array is the testing of tens of thousands of markers simultaneously, providing a comprehensive genome-wide view of the amount of DNA present. In one case, a mosaic trisomy was detected in a sample of 70% standard karyotype. A second case possessed a small deletion (1 Mbp) that was validated in a parent-offspring pair. More significantly, suitable material for cell culture (and thus FISH) was not available from two cases, yet a diagnosis was still possible from DNA extracted from blood and hair. Although the discontinuation of the Illumina arrays signifies the temporary loss of this technique for the horse (as the Affymetrix array cannot be used in the same manner, although the creation of a custom array is a possibility), this research demonstrates proof of concept for other domestic species.

Chapter Three presents the use of a *de novo* transcriptome assembly to describe a unique tissue and produce valuable annotation for future research. As most gene expression studies in the horse utilize quantitative PCR to assess known targets, proper annotation of gene models is crucial. Similarly, gene expression is generally tissue-specific, so it is critical to know which isoforms are appropriate to measure. The *de novo* assembly of lamellar tissue meets both of these needs, and thus provides an invaluable resource for laminitis research. Additionally, SNPs were called for each of the three sequenced horses and submitted to public databases, increasing the existing database of genomic variation. Currently, there is an effort in the equine genetics community to produce a resource of normal variation, as the NCBI dbSNP resource is generally under-utilized by livestock researchers. Limited funding for non-traditional research species often leads to smaller sample sizes in association studies. A variation catalog like this will allow rapid filtering of putative causal mutations from candidate loci, especially in next-generation sequencing datasets. While whole exome

sequencing provides a more economical view of coding regions, it requires proper gene annotation in order to design a capture array. RNA-seq is able to bypass this limitation, and although not all genes will be expressed, the enrichment for coding regions is still beneficial for future work.

Chapter Four demonstrates this concept by using RNA-seq to assist in mapping a trait. Linkage mapping previously identified a 16 Mbp region on a single chromosome, so a full genome-wide association study was unnecessary, but genome capture for variant identification was not practical. Also, as shown by the failure of targeted resequencing to detect the mutation responsible for *LP*, this method can be limited by the reference genome sequence (Bellone *et al*. 2010). The use of RNA-seq in a small number of samples enabled screening genes for variation only observed in the single case sequence, which aided in the development of two fine-mapping SNP arrays. A single SNP was strongly associated with the *PATN1* phenotype, and although it may not represent the casual variant, it can immediately be utilized by breeders for testing purposes. In several breeds, *PATN1* produces a highly valued coat color. However, this color is only seen in individuals that also possess at least one *LP* allele. With the ability to use this DNA based test to screen non-*LP* horses for the valued *PATN1* allele, breeders can identify otherwise overlooked offspring that may have value in improving the spotting patterns produced in their programs. As exome sequencing is not currently in common use (with only two publications in domestic animals indexed in PubMed), RNA-seq represents a viable alternative when capture arrays are not available (Ahonen *et al*. 2013, Cosart *et al*. 2011).

In non-traditional and non-model organisms, available genetic resources are frequently poor or non-existent. The domestic horse is fortunate to have such a wealth of tools available. However, amongst the other equids, these may not be sufficient, as there has been rapid karyotypic evolution throughout the family (Trifonov *et al*. 2008).

With the cost of sequencing sharply falling (Illumina having announced they can produce a human genome for $1000 in January of 2014), genomics approaches are becoming a possibility for research in all species. Similar to what has been discussed in this dissertation, RNA-seq has been used in other species for variant discovery and annotation generation (Gao *et al*. 2013, Gui *et al*. 2013, Liu *et al*. 2013, Christensen and Anistoroaei 2014). Variant discovery from RNA-seq has also been used for population genomics, generating diversity statistics from samples of pooled individuals without the need of a reference genome (Gayral *et al*. 2013, Zavodna *et al*. 2013). Next generation sequencing can also be applied to protect rare and endangered species. In one application, elicit material from endangered species was identified by DNA sequences found within traditional Chinese medicine products, indicating illegal poaching and violation of trade policies (Coghlan *et al*. 2012, Lammers *et al*. 2014).

This dissertation follows this trend, applying various technologies to provide resources for all equine professionals, from clinicians to researchers to breeders. Future research will hopefully incorporate these outputs to provide further tools for the health of the horse, whether it be through genetic testing or the development of novel biomarkers or drugs for laminitis. Although the immediate benefit is to the domestic horse, these same methods can be applied to aide other species.

REFERENCES

Ahonen SJ, Arumilli M, Lohi H (2013) A CNGB1 frameshift mutation in Papillon and Phalène dogs with progressive retinal atrophy. PLoS One 8: e72122.

Bellone RR, Forsyth G, Leeb T, Archer S, Sigurdsson S, Imsland F, Mauceli E, Engensteiner M, Bailey E, Sandmeyer L, Grahn B, Lindblad-Toh K, Wade CM (2010) Fine-mapping and mutation analysis of TRPM1: a candidate gene for leopard complex (LP) spotting and congenital stationary night blindness in horses. Brief Funct Genomics 9: 193-207.

Bellone RR, Holl H, Setaluri V, Devi S, Maddodi N, Archer S, Sandmeyer L, Ludwig A, Foerster D, Pruvost M, Reissmann M, Bortfeldt R, Adelson DL, Lim SL, Nelson J, Haase B, Engensteiner M, Leeb T, Forsyth G, Mienaltowski MJ, Mahadevan P, Hofreiter M, Paijmans JL, Gonzalez-Fortes G, Grahn B, Brooks SA (2013) Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. PLoS One 8: e78280.

Christensen K, Anistoroaei R (2014) An American mink (Neovison vison) transcriptome. Anim Genet 45: 301-303.

Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genet 8: e1002657.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC Genomics 12: 347.

Gao X, Han J, Lu Z, Li Y, He C (2013) De novo assembly and characterization of spotted seal Phoca largha transcriptome using Illumina paired-end sequencing. Comp Biochem Physiol Part D Genomics Proteomics 8: 103-110.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, Belkhir K, Cahais V, Loire E, Bernard A, Galtier N (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet 9: e1003457.

Gui D, Jia K, Xia J, Yang L, Chen J, Wu Y, Yi M (2013) De novo assembly of the Indo-Pacific humpback dolphin leucocyte transcriptome to identify putative genes involved in the aquatic adaptation and immune response. PLoS One 8: e72417.

Henderson LB, Applegate CD, Wohler E, Sheridan MB, Hoover-Fong J, Batista DA (2014) The impact of chromosomal microarray on clinical management: a retrospective analysis. Genet Med 2014 Mar 13.

Lammers Y, Peelen T, Vos RA, Gravendeel B (2014) The HTS barcode checker pipeline, a tool for automated detection of illegally traded species from high-throughput sequencing data. BMC Bioinformatics 15: 44.

Liu H, Wang T, Wang J, Quan F, Zhang Y (2013) Characterization of Liaoning cashmere goat transcriptome: sequencing, de novo assembly, functional annotation and comparative analysis. PLoS One 8: e77062.

Metallinos DL, Bowling AT, Rine J (1998) A missense mutation in the endothelin-B receptor gene is associated with Lethal White Foal Syndrome: an equine version of Hirschsprung disease. Mamm Genome 9: 426-431.

Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59: 5-15.

Rudolph JA, Spier SJ, Byrns G, Rojas CV, Bernoco D, Hoffman EP (1992) Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding. Nat Genet 2: 144-147.

Santschi EM, Purdy AK, Valberg SJ, Vrotsos PD, Kaese H, Mickelson JR (1998) Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses. Mamm Genome 9: 306-309.

Shin EK, Perryman LE, Meek K (1997) A kinase-negative mutation of DNA-PK(CS) in equine SCID results in defective coding and signal joint formation. J Immunol 158: 3565-3569.

Towers RE, Murgiano L, Millar DS, Glen E, Topf A, Jagannathan V, Drögemüller C, Goodship JA, Clarke AJ, Leeb T (2013) A nonsense mutation in the IKBKG gene in mares with incontinentia pigmenti. PLoS One 8: e81625.

Trifonov VA, Stanyon R, Nesterenko AI, Fu B, Perelman PL, O'Brien PC, Stone G, Rubtsova NV, Houck ML, Robinson TJ, Ferguson-Smith MA, Dobigny G, Graphodatsky AS, Yang F (2008) Multidirectional cross-species painting illuminates the history of karyotypic evolution in Perissodactyla. Chromosome Res 16: 89-107.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole

Genome Assembly Team, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867.

Yang GC, Croaker D, Zhang AL, Manglick P, Cartmill T, Cass D (1998) A dinucleotide mutation in the endothelin-B receptor gene is associated with lethal white foal syndrome (LWFS); a horse variant of Hirschsprung disease. Hum Mol Genet 7: 1047-1052.

Zavodna M, Grueber CE, Gemmell NJ (2013) Parallel tagged next-generation sequencing on pooled samples - a new approach for population genetics in ecology and conservation. PLoS One 8: e61471.